Final Report of Internship Program

On

"PREDICT BLOOD DONATION"

MEDTOUREASY, NEW DELHI

28th January, 2021

By: Abhishek Gaurav

ACKNOWLEDGEMENT

The internship opportunity that I had with MedTourEasy was a great change for learning and understanding the intricacies in the field of Data Analytics in Data Science; and also, for personal as well as professional development. I am very obliged for having a chance to interact with so many professionals who guided me throughout the internship project and made it a great learning curve for me.

Firstly, I express my deepest gratitude and special thanks to the Training Head of MedTourEasy, Mr. Ankit Hasija who gave me an opportunity to carry out my internship at their esteemed organization. Also, I express my thanks to him for making me understand the details of the Data Analytics profile and training me in the same so that I can carry out the project properly and with maximum client satisfaction and also for spearing his valuable time in spite of his busy schedule.

I would also like to thank the team of MedTourEasy who made the working environment productive and very conducive.

ABSTRACT

Forecasting blood supply is a serious and recurrent problems that a blood collection manager deals with. The title of the project is "**Predict Blood Donation**". A blood transfusion is a way of adding blood to the body after an injury or illness. Blood transfusion saves lives - from replacing lost blood during major surgery or a serious injury to treating various illnesses and blood disorders. Ensuring that there's enough blood in supply whenever needed is a serious challenge for the health professionals. The demand for blood fluctuates throughout the year. As one prominent example, blood donations slow down during busy holiday seasons. An accurate forecast for the future supply of blood allows for an appropriate action to be taken ahead of time and therefore saving more lives.

It is generally seen that blood supply decreases during winters and peoples who are on travel or errands are very less likely to undergo blood donations. The project undergoes a development pipeline automated by TPOT auto-ML library which is discussed in detail further in this report.

TABLE OF CONTENTS

INTRODUCTION

1.1 About the Company

MedTourEasy, an online medical tourism marketplace, provides you the informational resources needed to evaluate your global options. It helps you find the right healthcare solution based on your specific health needs, affordable care, while meeting the quality standards that you expect to have in healthcare.

MedTourEasy, a global healthcare company, provides you the informational resources needed to evaluate your global options. It helps you find the right healthcare solution based on specific health needs, affordable care while meeting the quality standards that you expect to have in healthcare.

MedTourEasy improves access to healthcare for people everywhere. It is an easy to use platform and service that helps patients to get medical second opinions and to schedule affordable, high-quality medical treatment abroad.

MedTourEasy commitment to quality and transparency in healthcare is core to our mission. At MedTourEasy, we integrate the same three factors that physicians themselves agree are most important when selecting or referring a healthcare provider (patient satisfaction, experience match, and the quality of the hospital where a physician provides care.

MedTourEasy aspires to be the leader in making information on physicians and hospitals more accessible and transparent. The purpose is to give people the confidence to make the right healthcare decisions.

MedTourEasy's mission is to provide access to quality healthcare for everyone, regardless of location, time frame, or budget. Patients can connect with internationally-accredited clinics and hospitals.

1.2 About the Project

The title of the project is “Predict Blood Donations”. Forecasting blood supply is a serious and recurrent problems that a blood collection manager deals with. A blood transfusion is a way of adding blood to the body after an injury or illness. Blood transfusion saves lives - from replacing lost blood during major surgery or a serious injury to treating various illnesses and blood disorders. Ensuring that there's enough blood in supply whenever needed is a serious challenge for the health professionals. The demand for blood fluctuates throughout the year. As one prominent example, blood donations slow down during busy holiday seasons. An accurate forecast for the future supply of blood allows for an appropriate action to be taken ahead of time and therefore saving more lives.

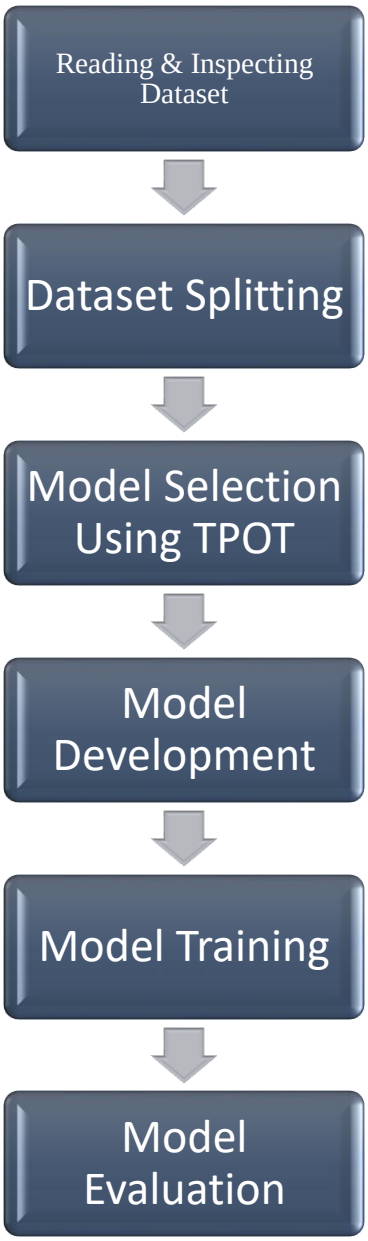
The project involves working on blood transfusion dataset. It includes the following steps:-

- Loading the blood transfusion dataset
- Inspecting the dataframe
- Specifying features and target columns
- Splitting dataset into training and testing dataset
- Selecting model using TPOT
- Model Building
- Model Training
- Model Evaluation

METHODOLOGY

2.1 Flow of the Project

The flow of the project can be understood via the following diagram:-



2.2 Language and Platform Used

Language: **Python**

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985-1990. Like Perl, Python source code is also available under the GNU General Public License (GPL).

Fig. Creator of Python

Following are important characteristics of Python Programming:

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- It supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

Machine Learning:

According to Arthur Samuel, "Machine Learning algorithms enable the computers to learn from data, and even improve themselves, without being explicitly programmed."

Machine learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available.

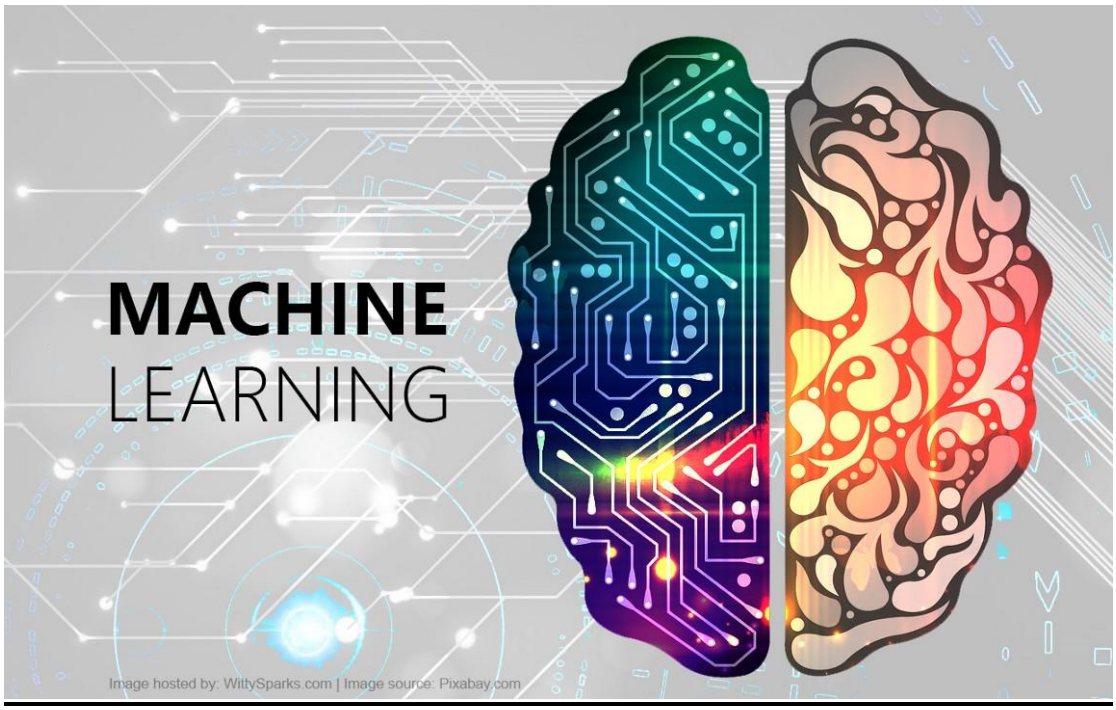


Fig. Machine Learning

Types of Machine Learning:

Machine Learning can be classified into three categories of algorithms:-

i. Supervised Learning
ii. Unsupervised Learning

iii. Reinforcement Learning

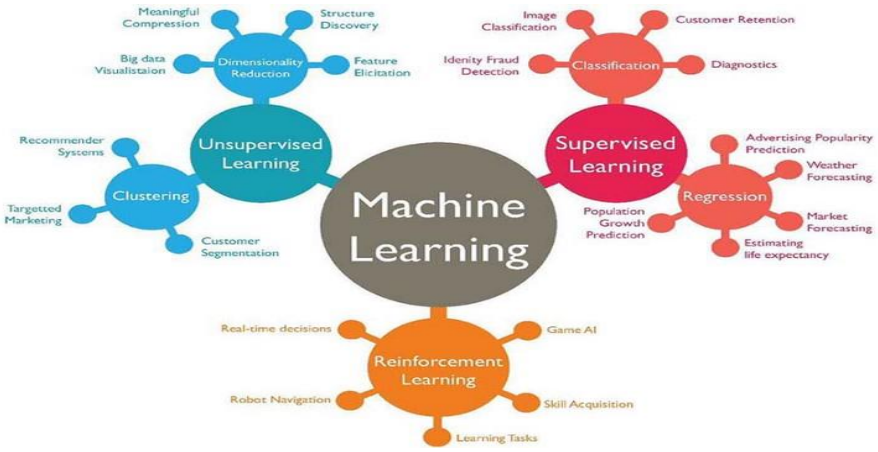


Fig. Types of Machine Learning

Platform: **Jupyter Notebook**

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

IMPLEMENTATION

3.1 Dataset Description

The dataset, 'transfusion.csv', obtained from the Machine Learning Repository, consists of a random sample of 748 donors. Our dataset is from a mobile blood donation vehicle in Taiwan. The Blood Transfusion Service Center drives to different universities and collects blood as part of a blood drive. We want to predict whether or not a donor will give blood the next time the vehicle comes to campus. It is structured according to RFMTC marketing model (a variation of RFM).

RFM stands for Recency, Frequency and Monetary Value and it is commonly used in marketing for identifying your best customers. In our case, our customers are blood donors.

RFMTC is a variation of the RFM model. Below is a description of what each column means in our dataset:

- R (Recency - months since the last donation)
- F (Frequency - total number of donation)
- M (Monetary - total blood donated in c.c.)
- T (Time - months since the first donation)
- a binary variable representing whether he/she donated blood in March 2007 (1 stands for donating blood; 0 stands for not donating blood)

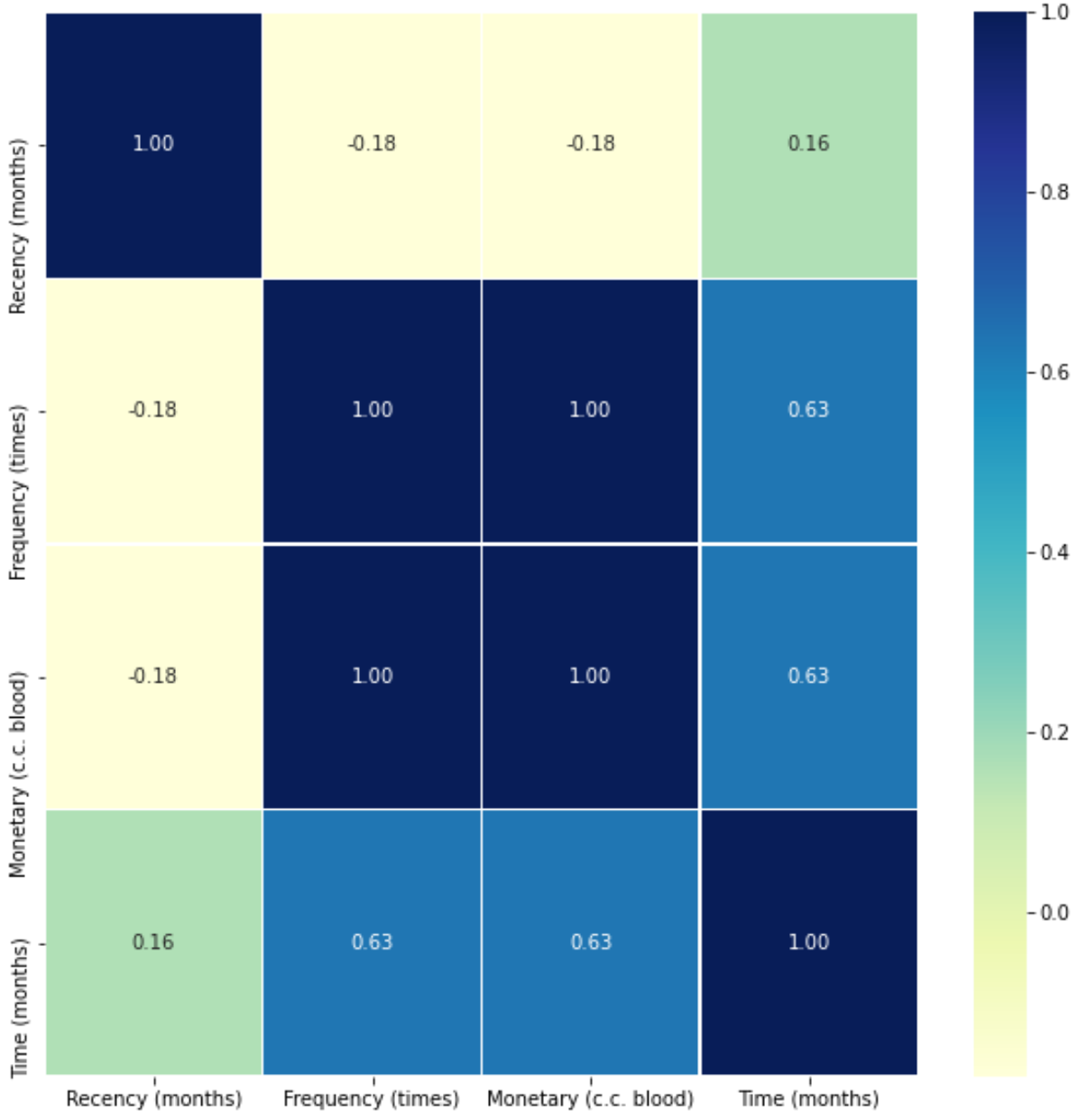
3.2 Statistical Insights of Dataset

Every dataset contains multiple hidden information which can be seen by performing statistical analysis on it. The insights found by performing statistical analysis on our dataset are the following:-

- Every column in our dataset is of numeric type.
- After specifying the features and target column, it is found that the distribution of target class in target columns are as:
  - **0 : 0.762**
  - **1 : 0.238**
- Due to the uneven distribution of both classes, splitting of dataset is performed using the following parameters:-
  - Test size : **0.25**

- Random State : **42**
- Stratify : Target column

- The correlation matrix describing the correlation between different features through a heatmap is shown below:-



3.3 Model Selection and Development

To select a perfect algorithm and whole development pipeline is quite a tedious and time taking task. Therefore, we use Auto ML which defines the best pipeline and best machine learning algorithm with the best parameters.

Here, I have used **TPOT**.

It stands for "**The Tree-Based Pipeline Optimization Tool**". TPOT is a Python Automated Machine Learning tool that optimizes machine learning pipelines using genetic programming.

Fig. Logo of TPOT

TPOT automates the most tedious part of machine learning by intelligently exploring thousands of possible pipelines to find the best one for our data. TPOT is built on top of scikit-learn, so all of the code it generates should look familiar.

TPOT picked "*Logistic Regression"* as the best model for our dataset with no pre-processing steps, giving us the AUC score of 0.7850.

Therefore, I have used Logistic regression algorithm for model development.

3.4 Model Training & Evaluation

One of the assumptions for linear regression models is that the data and the features we are giving it are related in a linear fashion, or can be measured with a linear distance metric. If a feature in our dataset has a high variance that's an order of magnitude or greater than the other features, this could impact the model's ability to learn from other features in the dataset.

Correcting for high variance is called normalization. It is one of the possible transformations we do before training a model.

Monetary (c.c. blood)'s variance was very high in comparison to any other column in the dataset. This means that, unless accounted for, this feature might get more weight by the model (i.e., be seen as more important) than any other feature. One way to correct for high variance was to use log normalization.

Finally, the logistic regression model is trained with the following parameters:

- Solver = liblinear
- Random State = 42

The model evaluation is performed by checking AUC Score of the model which is **0.7891**.

The code for the project can be viewed at:-

https://drive.google.com/drive/folders/1iLb0w5iatMaUrQXdl3_mDR3kIJiyHBHJ?usp=sharing

CONCLUSION & FUTURE SCOPE

The demand for blood fluctuates throughout the year. As one prominent example, blood donations slow down during busy holiday seasons. An accurate forecast for the future supply of blood allows for an appropriate action to be taken ahead of time and therefore saving more lives.

In this project, we explored automatic model selection using TPOT and AUC score which was **0.7850**. This is better than simply choosing '0' all the time (the target incidence suggests that such a model would have 76% success rate). We then log normalized our training data and improved the AUC score by 0.5%. In the field of machine learning, even small improvements in accuracy can be important, depending on the purpose.

Another benefit of using logistic regression model is that it is interpretable. We can analyze how much of the variance in the response variable (target) can be explained by other variables in our dataset.

Pre-donation information and counselling are linked to the process of donor selection in which each individual's suitability to donate is carefully assessed against a set of criteria related to their medical history.

Pre-donation information is an important first step in informing and educating donors about the blood donation process, including donor selection criteria and deferral or self-deferral, blood screening for TTI, blood grouping, counselling and referral. This will enable individuals who may be unsuitable to donate blood to self-defer without going through the blood donation process.

REFERENCES

The following websites have been referred for input data and statistics:-

- https://www.webmd.com/a-to-z-guides/blood-transfusion-what-to-know#1
- https://www.kjrh.com/news/local-news/red-cross-in-blood-donation-crisis
- https://www.ncbi.nlm.nih.gov/books/NBK310569/
- http://epistasislab.github.io/tpot/

The following websites have been referred for coding part:-

- https://www.python.org/
- https://github.com/perborgen/LogisticRegression
- http://epistasislab.github.io/tpot/