PREDICTION OF BLOOD DONATION

Final report work of the Traineeship Program in

MedTourEasy

Submitted by :

Abhilasha Chatterjee

November 2024

ACKNOWLEDGEMENT

The traineeship with MedTourEasy was a great opportunity for me to learn and understand the intricacies of the subject of Data Processing and Analysis; 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 traineeship project and made it a great learning curve for me.

Firstly, I would like to express my deepest gratitude and special thanks to the Training & Development Team of MedTourEasy who gave me an opportunity to carry out my traineeship at their esteemed organization.

Also, I am overwhelmed in all humbleness and gratefulness to the team for making me understand the details of the profile and training me for the same in carrying out the project properly and with maximum client satisfaction.

Thanking You

Abhilasha Chatterjee

ABSTRACT

Blood donation is a vital part of worldwide healthcare. It relates to It relates to blood transfusion as a life-sustaining and life-saving procedure as well as a form of therapeutic phlebotomy as a primary medical intervention. Over one hundred million units of blood are donated each year throughout the world.

This research study is focused on the analysis and prediction of blood donation of a Service Center in Taiwan who conducts donation drives across different university campuses indulging students to contribute for the society. We will see how the dataset performs if a predictive analysis using machine learning algorithm is done. This dataset is provided by the MedTourEasy company.

The objective of the dataset is to diagnostically predict whether or not a student will donate blood the next time a vehicle approaches the campus for blood collection based on certain diagnostic measurements included in the dataset. For analysis of the given dataset I have applied the TPOT model which selected Logistic Regression as the best model. I have done a performance measurement on the basis of that.

TABLE OF CONTENTS

INTRODUCTION

1.1 About the company

- MedTourEasy, a global healthcare company, provides us the informational sources needed to evaluate our global options. We can find the right healthcare solution based on specific health needs and affordable care while meeting the quality standards one expects. The company improves access to healthcare for people everywhere. It is an easy to use platform giving services to patients in getting quality medical treatment.

1.2 About the project

- A blood donation occurs when a person voluntarily has blood drawn and used for transfusions or is made into bio-pharmaceutical medications by a process called fractionation i.e. separation of whole blood components. A donation may be of whole blood, or of specific components directly.
- Today in the developed world, most blood donors are unpaid volunteers who donate blood for a community supply. In some countries, established supplies are limited and donors usually give blood when family or friends need a transfusion. Many donors donate for several reasons, such as a form of charity, general awareness regarding the demand for blood, increased confidence in oneself or helping a personal friend or relative, etc. Despite the many reasons that people donate, not enough potential donors actively donate.
- This research study is based on the dataset collected from a donor database of Blood Transfusion Service Center. The dataset is from a mobile blood donation vehicle in Taiwan, China. The Service Center has vehicles that are driven to different university campuses for collection of blood as part of the drive.

RESEARCH OBJECTIVES

The project focuses on developing a prediction model by selecting variables from the array of indicators categorized under different heads.

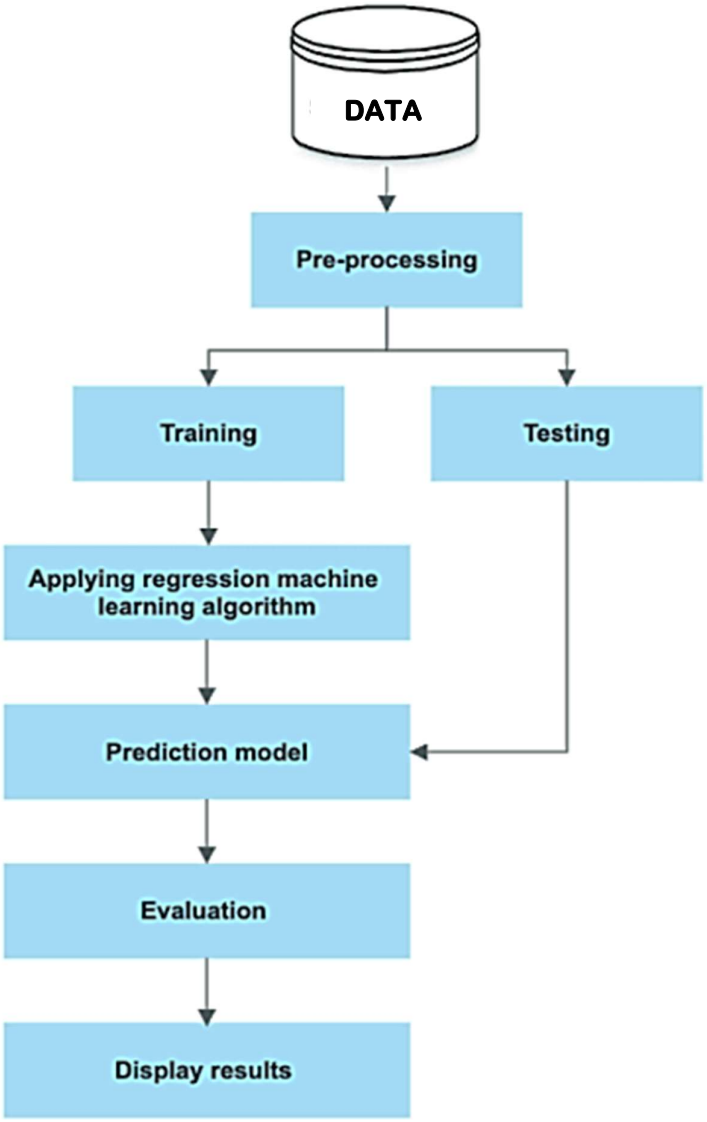
- The focus here is to apply a logistic regression model for the prediction.
- Also to create interactive visuals that will help in analyzing the situation and drawing conclusions accordingly.

The two models' accuracy of prediction is then compared against each other to address the differences in the model.

METHODOLOGY

2.1 Flow of the Project

The project followed the following steps to accomplish the desired objectives. Each step has been followed as it is to achieve the model required along with the results.



2.2 Language and Platform Used

- Language : Python

  Python is a very popular, general-purpose, interpreted, interactive, object-oriented, and high-level programming language. Its design philosophy emphasizes code readability with the use of significant indentation.
  Guido van Rossum began working on Python in the late 1980s as a successor to the ABC programming language.
  Python is also dynamically typed as it supports multiple programming paradigms, including structured, object-oriented and functional programming. It is more often described as a "batteries included" language mostly due to its comprehensive standard library. Python consistently ranks as one of the most popular programming languages, and has gained widespread use in the machine learning community.
  Its core features are :

  - Easy to learn and use
  - Versatility in various domains
  - Open source and interpreted language
  - Wide range of libraries and frameworks
  - Dynamic memory allocation

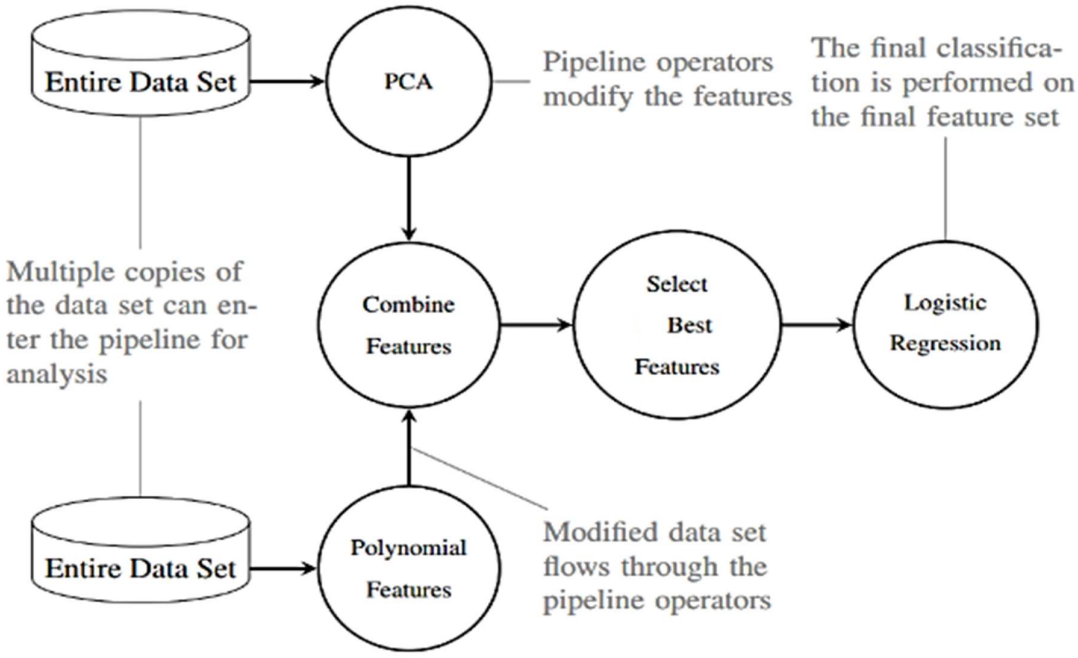
- Platform : Jupyter Notebook

  Jupyter is a project to develop open-source software, open standards, and services for interactive computing across multiple programming languages.
  Jupyter Notebook is a web-based interactive computational environment for creating notebook documents. It is built using several open-source libraries. A Jupyter Notebook application is a browser-based REPL containing an ordered list of input/output cells which can contain code, text, mathematics, plots and rich media.
  A Jupyter Notebook document is a JSON file, following a versioned schema, usually ending with the ".ipynb" extension. The main parts of the Jupyter Notebooks are: Metadata, Notebook format and list of cells.
  The key features are :

  - Integration of Markdown-formatted text
  - Rich outputs like tables and charts are displayed
  - Flexibility in terms of language switching.
  - Adaptability and extensibility via add-ons.
  - Quick feedback and live code execution.
  - Widely employed in scientific research and education.

- Package : Tpot

  TPOT i.e. Tree-based Pipeline Optimization is a Python Automated Machine Learning tool that optimizes machine learning pipelines using genetic programming. TPOT will automate 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 are similar to that.



IMPLEMENTATION

3.1 Defining the problem statement

The first step is to collect the data to understand the goals of the project which are to be achieved based on the problem statement defined which has to be referred for the project development.

In this case, we need to predict whether the university students would be willing to donate blood the next time a Service Center vehicle approaches for blood collection.

3.2 Data Collection and Importing

- Data Collection is a systematic approach for gathering and measuring information from a variety of sources to obtain a complete and accurate picture of an interest area. This helps an individual or organization to address specific questions, determine outcomes and forecast future trends or patterns.
- Data Importing is the uploading of the required data into the coding environment from internal sources (computer) or external sources (online websites or data repositories). This data can be manipulated, aggregated and filtered as per the needs and requirements of the project.

DATASET

	Recency (months)	Frequency (times)	Monetary (c.c. blood)	Time (months)	whether he/she donated blood in March 2007
0	2	50	12500	98	1
1	0	13	3250	28	1
2	1	16	4000	35	1
3	2	20	5000	45	1
4	1	24	6000	77	0

3.3 Data Cleaning

Data is the most imperative aspect of analysis and is required everywhere.
But many a times, the data may be incomplete, inconsistent or contain missing values when it comes to the real world. If data is corrupted, then it will hamper the process by giving inaccurate results. Therefore, data cleaning is considered a foundational element of the basic data science.
It is the process by which the incorrect, incomplete, inaccurate, irrelevant or missing part of the data is identified and then modified, replaced or deleted as needed.

```
<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
```

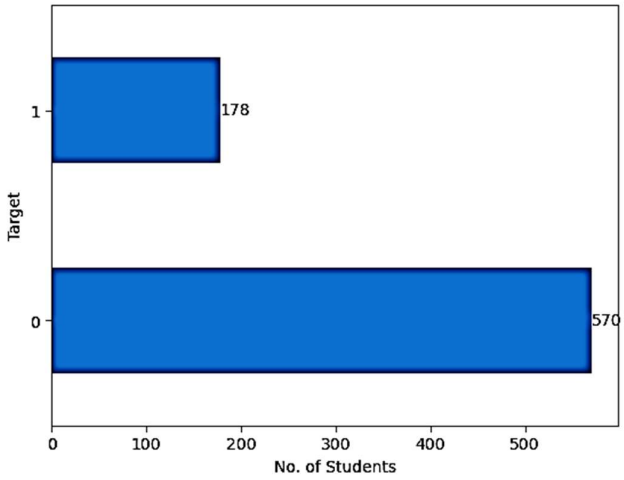
From the extracted information, we can see that the dataset is not having any missing values. We could have used various functions to check for null values or incorrect values but there seems to be no such mess in the data. Hence our dataset is clean and we can proceed further with the analysis.

3.4 Data Visualization

The process of presenting data in a graphical or pictorial formal is visualization. It allows decision makers to see visually represented outputs of the analysis done, hence helping them in grasping difficult concepts easily and identify new patterns. Visuals can be in the form of tables, charts or graphs as it is easier to visualize large amounts of complex data in these forms rather than on spreadsheets or reports. Data visualization helps us in many ways:

- Identify areas that need attention or improvement
- Clarify factors influencing customer behavior
- Predict sales volume, etc.

In this case, we have a “Target” for us that is whether students will donate blood or not. This is a case of binary classifier, meaning it has only two possible outcomes.
0 : not donating blood
1 : donating blood



3.5 Data Analysis

- Model Selection
  - Splitting the dataset into training and testing data

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(transfusion.drop(columns='Target'),
    transfusion.Target, test_size=0.25, random_state=42, stratify=y)
```

```
X_train.shape
```

```
(561, 4)
```

```
X_test.shape
```

```
(187, 4)
```

  - Using the TPOT Model

TPOT is a Python Automated Machine Learning tool that optimizes machine learning pipelines using genetic programming.
This model is efficient as it automatically explores all the possible pipeline to suggest the best one for our dataset.

```
Best pipeline: LogisticRegression(input_matrix, C=0.1, dual=False, penalty=l2)
```

```
TPOTClassifier
TPOTClassifier(config_dict='TPOT light', disable_update_check=True,
               generations=5, population_size=20, random_state=42,
               scoring='roc_auc', verbosity=2)
```

As Tpot Classifier suggests a regression model will be the best fit. So we are next going to develop our Logistic Regression Model for the analysis.

- Log Normalization

It is a method of standardizing data that can be useful in case of some variables in a dataset having high variance.
Logarithmic transformation is applied to our values, transforming them into a scale that approximates normality – an assumption usually made.

Checking the variance :

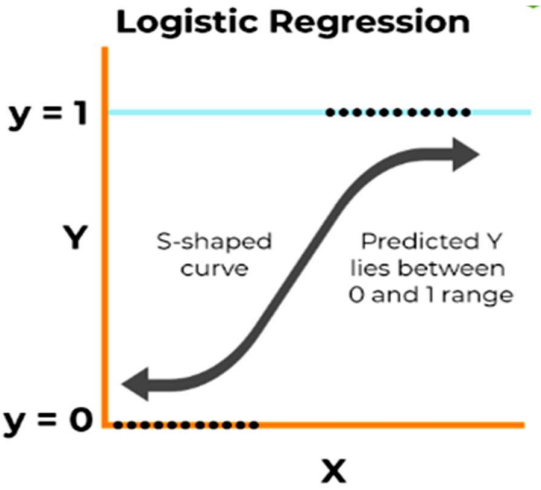
```
Recency (months)              66.929
Frequency (times)             33.830
Monetary (c.c. blood)    2114363.700
Time (months)                611.147
dtype: float64
```

Normalizing the data :

```
Recency (months)          66.929
Frequency (times)         33.830
Time (months)            611.147
Monetary(c.c. blood)       0.835
dtype: float64
```

- Model Training

Development of Logistic Regression Model



Logistic regression analysis is conducted when the dependent variable is dichotomous (binary). It is used in predictive analysis like other regression models. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables.

A logistic regression analysis is undertaken when we have a binary dependent variable in our model. Moreover, the effects of outliers are minimized in logistic regression. Linear regression models are highly affected by outliers as the best fit line shifts to minimize the distance between the predicted value and actual value.

Assumptions of the model :

1. The outcome is binary
   - The outcome of a logistic regression model is binary, i.e., can only take two values like bankrupt and non-bankrupt.
2. The logistic of the outcome and independent variable have a linear relationship
   - The outcome of the model and the explanatory variable should have a linear relationship.
3. No homoskedasticity
   - The logistic model doesn't require an assumption of homoskedasticity, i.e., when the error term (that is, the "noise" or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables.
4. No severe multicollinearity
   - The logistic regression model necessitates that there should not be high multi-collinearity between the independent variables.

Function of the model :

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

$f(x)$ = output of the function
L = the curve's maximum value
k = logistic growth rate or steepness of the curve
x_0 = the x value of the sigmoid midpoint
x = real number

Logistic Regression is the most common method used for binary classification problems. The function used in Logistic regression is the logistic function which is an S-shaped curve that can take any number and map it into a value between 0 and 1.
Input values are combined linearly using coefficient values to predict an output value.

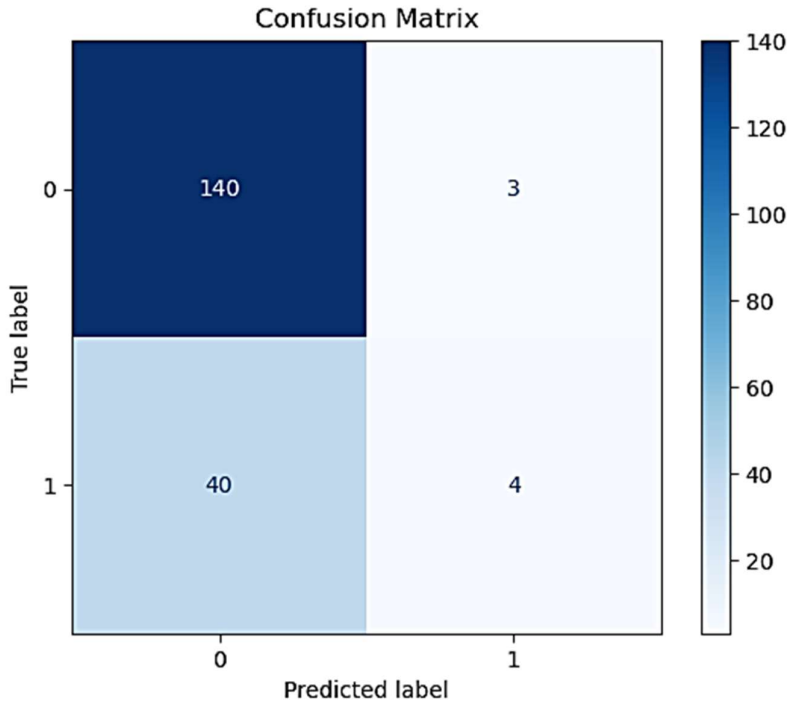
OBSERVATION

This research dataset is divided into two parts, two-thirds of the data is used as a training set, and one-third of the dataset is defined as a testing set to evaluate the performance of several variables.
Since our "Target" variable showed that 0s appeared 76% of the time. We want to make the prediction model keeping the structure same for training and testing data.

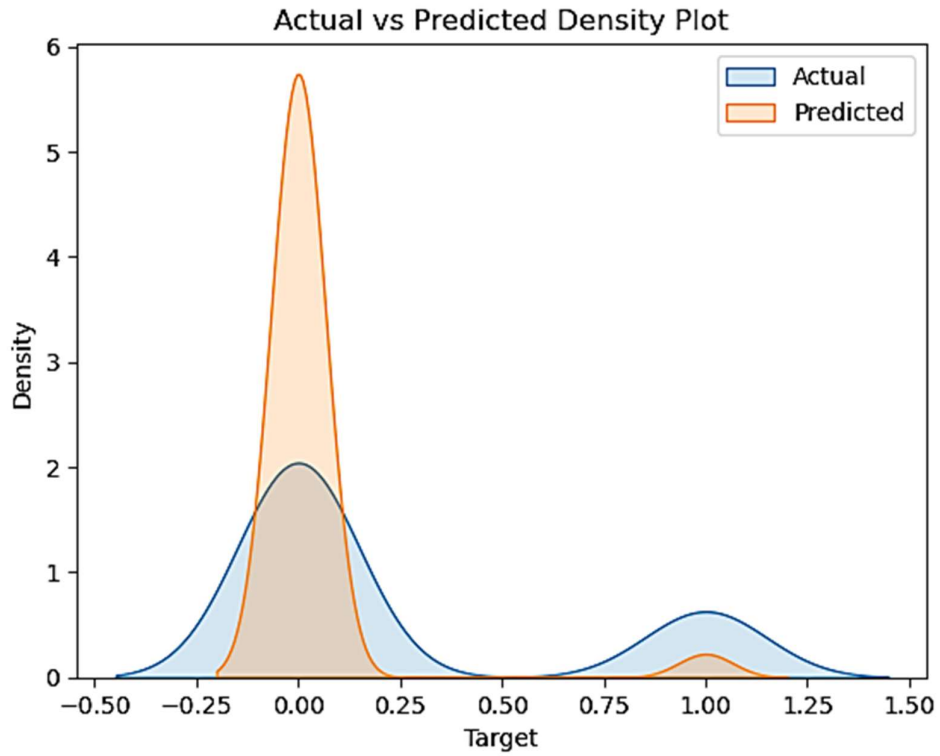
```
Target
0    0.762
1    0.238
Name: proportion, dtype: float64
```

To see if the model gives good results at identifying 1's and 0's, we check the confusion matrix. A confusion matrix is a table that is used to define the performance of a classification algorithm by visualizing and summarizing it.

```
Confusion Matrix:
 [[140   3]
 [ 40   4]]
```

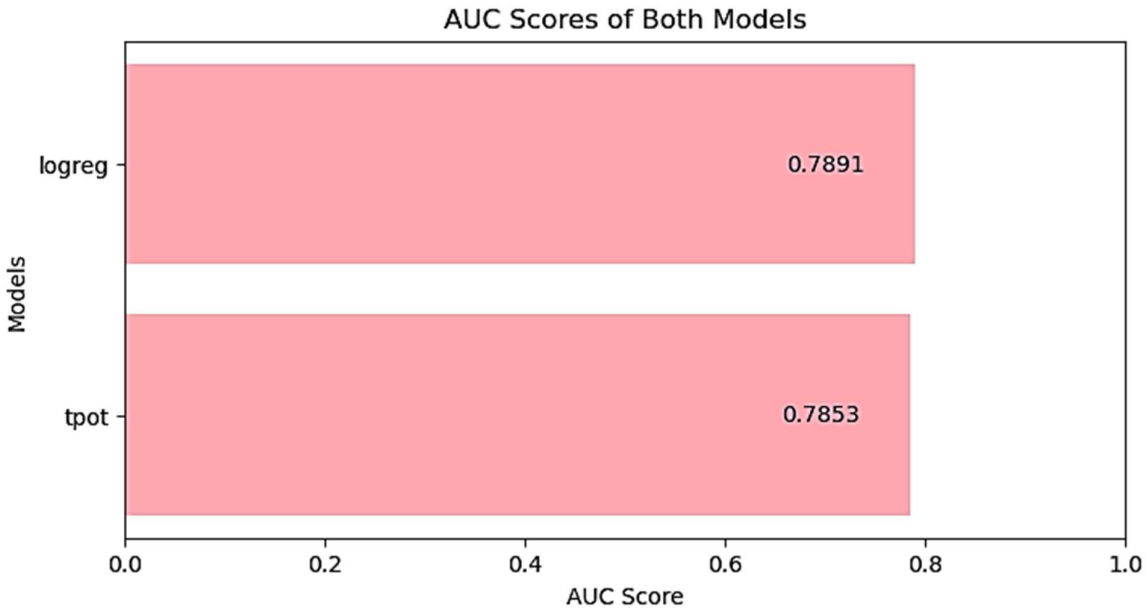


Next we check for the performance of the dataset on a density plot.
A density plot is a graph that shows the distribution of a numeric variable as a smooth curve. It uses a kernel density estimate to show the probability density function of the variable. It is a smoothed version of a histogram that can help reduce noise of data.



We calculated the AUC score (Area Under the Curve) is a performance metric used primarily in classification problems, especially binary classification. It evaluates how well a model distinguishes between classes, typically using the ROC (Receiver Operating Characteristic) curve or the Precision-Recall (PR) curve.

```
Sorted models by AUC score (highest to lowest):
logreg: 0.7891
tpot: 0.7853
```



We made a Classification report to check the accuracy, which overall tells us how are model is correct. We see precision, recall and F1 score, which gives us insights into how well our model performed at correctly identifying different classes.

	precision	recall	f1-score	support
0	0.78	0.98	0.87	143
1	0.57	0.09	0.16	44
accuracy			0.77	187
macro avg	0.67	0.53	0.51	187
weighted avg	0.73	0.77	0.70	187

The Logistic Regression model has an accuracy of 77%.

CONCLUSION

Usually the donation of blood depends upon the following factors:

1. Age and weight
2. Hemoglobin count
3. Medical condition
4. Blood pressure

From the exploratory data analysis performed, this can be concluded that there is a 77% probability that the students donate blood the next time a Service Center vehicle approaches their campus.

According to the TPOT model, Logistic Regression is the best algorithm out of all the models with 77% accuracy. The accuracy of the model can be improved by increasing the size of the dataset.
The dataset used here is very small and had only 748 rows.

BIBLIOGRAPHY

The Link provided by MedTourEasy for the dataset and notebook :

- https://drive.google.com/file/d/1S2o3wEAfEPha06ECh6kirwUijcCq54nY/view?usp=sharing

Personal References

- https://en.wikipedia.org/wiki/Blood_donation
- https://epistasislab.github.io/tpot/
- https://campus.datacamp.com/courses/preprocessing-for-machine-learning-in-python/standardizing-data?ex=4