Final Report of Internship Program

On

“PREDICT BLOOD DONATION”

MEDTOUREASY, NEW DELHI

28th June, 2023

By: Shweta Kumavat

ACKNOWLEDGEMENT

The internship I had with MedTourEasy was a fantastic opportunity for me to learn and understand the ins and outs of Data Analytics in Data Science. It was a time of significant personal and professional growth for me. I'm extremely grateful to have had the chance to interact with many professionals who guided me throughout the internship project and made it a valuable learning experience.

Firstly, I want to express my deepest gratitude and give special thanks to Mr. Ankit Hasija, the Training Head of MedTourEasy, who gave me this internship opportunity. I also want to thank him for taking the time to explain the details of the Data Analytics profile and providing the necessary training. Despite his busy schedule, he dedicated his valuable time to help me succeed.

I would also like to thank the entire team at MedTourEasy for creating a productive and supportive working environment.

Overall, I'm extremely grateful for the internship experience at MedTourEasy. It has helped me grow and gain valuable knowledge and skills that I will carry with me throughout my career.

ABSTRACT

The project "Predict Blood Donation" focuses on addressing the ongoing challenge faced by blood collection managers—forecasting blood supply. Blood transfusion plays a vital role in saving lives by replenishing blood lost during surgeries, injuries, and treating various illnesses. However, ensuring an adequate blood supply at all times is a significant task for healthcare professionals. The demand for blood fluctuates throughout the year, with slower donations during busy holiday seasons. Having an accurate forecast of future blood supply enables proactive measures to be taken, ultimately saving more lives.

One common observation is that blood supply tends to decrease during winter, as fewer people are likely to donate blood when they are traveling or occupied with other tasks. To address this issue, the project utilizes an automated development pipeline powered by the TPOT auto-ML library. This library's detailed discussion is provided later in the report.

TABLE OF CONTENTS

INTRODUCTION

1.1 About the Company

MedTourEasy is an online platform that helps you find the right healthcare solution worldwide, considering your specific health needs, affordability, and quality standards. It aims to improve access to healthcare for everyone by providing easy-to-use services such as medical second opinions and scheduling affordable, high-quality treatment abroad. MedTourEasy is committed to transparency and quality in healthcare, focusing on factors like patient satisfaction, experience match, and the quality of hospitals. The goal is to make information on physicians and hospitals more accessible, empowering people to make confident healthcare decisions. MedTourEasy connects patients with internationally-accredited clinics and hospitals, striving to provide access to quality healthcare regardless of location, time frame, or budget.

1.2 About the Project

The project "Predict Blood Donations" focuses on forecasting blood supply, which is a critical and recurring problem for blood collection managers. Blood transfusion plays a life-saving role by replenishing blood lost during surgeries, injuries, and treating various illnesses. Ensuring an adequate blood supply when needed is a significant challenge for healthcare professionals. The demand for blood fluctuates throughout the year, with donations slowing down during busy holiday seasons as one prominent example. Accurately forecasting future blood supply allows proactive actions to be taken, ultimately saving more lives.

In this project, we work with a blood transfusion dataset and perform the following steps:

1. Loading the blood transfusion dataset
2. Inspecting the dataset to understand its structure
3. Identifying the relevant features and target columns
4. Splitting the dataset into training and testing sets
5. Selecting the best model using TPOT (an automated machine learning library)
6. Building the selected model
7. Training the model using the training dataset
8. Evaluating the model's performance

By following these steps, we aim to develop a predictive model that can accurately forecast blood donations, aiding healthcare professionals in taking appropriate actions in advance to ensure sufficient blood supply and save more lives.

1.3 Objective and Deliverable

Objectives

- Develop a predictive model: The main objective is to build a model that can accurately forecast blood donations. This will help in anticipating the future supply of blood and enable proactive measures to be taken to meet the demand.
- Improve blood supply management: By accurately predicting blood donations, the project aims to enhance the management of blood supply. This will ensure that sufficient blood is available when needed, reducing the risk of shortages and saving more lives.
- Enhance decision-making: The project seeks to provide healthcare professionals and blood collection managers with valuable insights for making informed decisions related to blood supply and resource allocation. Accurate predictions will aid in planning and optimizing the allocation of resources.

Deliverables of the project

- Predictive model implementation: The project will deliver a functioning predictive model capable of forecasting blood donations based on historical data and relevant features. The model will be trained and ready for deployment.
- Performance evaluation: The model will be evaluated using appropriate metrics to assess its accuracy and reliability. This evaluation will provide insights into the model's effectiveness and its potential for real-world application.
- Documentation and reporting: Detailed documentation will be provided, outlining the methodology, data analysis techniques, model selection process, and implementation steps. A comprehensive report summarizing the findings, conclusions, and recommendations will be delivered.
- Recommendations for blood supply management: Based on the analysis and predictions, the project will provide recommendations to improve blood supply management, including strategies for addressing fluctuations in blood donations, identifying potential areas for improvement, and optimizing resource

allocation.

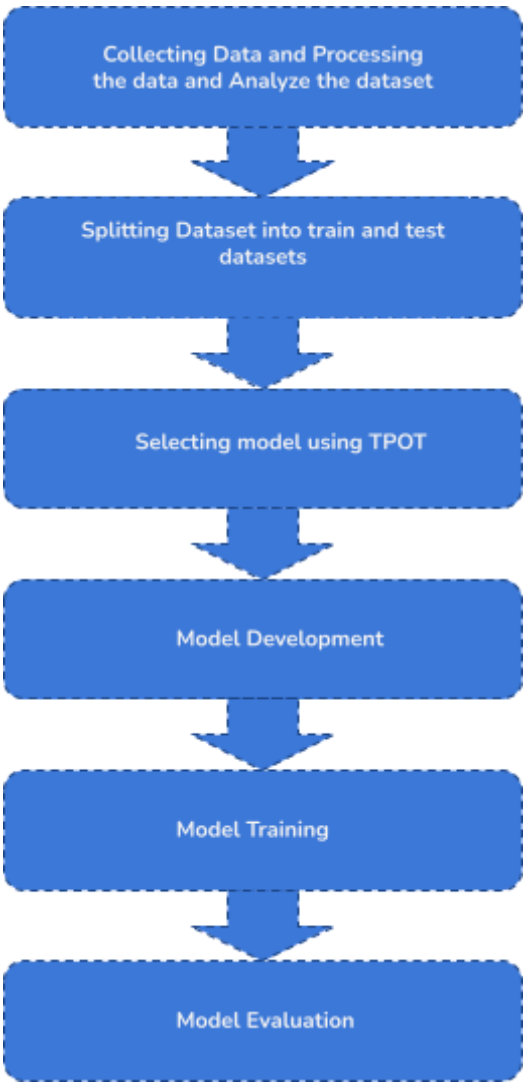
- Knowledge transfer and training: The project will include knowledge transfer sessions and training materials to ensure that stakeholders can effectively utilize and maintain the predictive model. This will enable the integration of the model into the existing blood supply management systems and processes.

By achieving these objectives and delivering the aforementioned items, the project aims to contribute towards more efficient and effective blood supply management, ultimately saving lives and improving healthcare outcomes.

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

The project "Predict Blood Donations" was implemented using the following languages and platforms:

- **Python**: Python programming language served as the foundation for the project. Python is widely acclaimed for its simplicity, readability, and extensive libraries that facilitate efficient data analysis, machine learning, and predictive modeling tasks.
- **Pandas**: Pandas library was employed for data manipulation and preprocessing. With its powerful data structures like DataFrames and versatile functions, Pandas provided a convenient way to handle, clean, and transform the blood donation dataset. It allowed for efficient data exploration, handling missing values, and feature engineering.
- **Scikit-learn**: Scikit-learn, a popular machine learning library in Python, played a crucial role in building and evaluating predictive models. It provided a comprehensive set of algorithms for classification, regression, and model selection. Scikit-learn also offered tools for data splitting, cross-validation, and performance evaluation metrics, enabling robust model development.
- **TPOT**: The Tree-based Pipeline Optimization Tool (TPOT) was utilized to automate the machine learning pipeline. TPOT employs genetic programming to search for the best combination of preprocessing steps and machine learning algorithms. By iteratively evaluating and evolving pipelines, TPOT helps in identifying optimal models for the given dataset, thereby saving time and effort in the model selection process.
- **Jupyter Notebook**: Jupyter Notebook served as the interactive development environment for coding, experimentation, and data exploration. It provided a web-based interface to write and execute Python code in a modular and well-documented manner. Jupyter Notebook allowed for seamless integration of code, visualizations, and explanations, facilitating clear and reproducible project workflows.

The combination of Python, Pandas, Scikit-learn, TPOT, and Jupyter Notebook offered a robust and efficient environment for data analysis, model development, and evaluation. These tools enabled seamless data manipulation, machine learning pipeline automation, and interactive code execution, ultimately contributing to the success of the "Predict Blood Donations" project.

IMPLEMENTATION

3.1 Dataset Description

The dataset we are using, named 'transfusion.csv', was obtained from the Machine Learning Repository. It consists of information about 748 blood donors who participated in a mobile blood donation event in Taiwan. The event was organized by the Blood Transfusion Service Center, which visits various universities to collect blood as part of their blood drive initiative.

Our objective is to predict whether a donor will give blood during the next visit of the blood donation vehicle to the university campus. To analyze the data, we have structured it according to the RFMTC marketing model, which is a variation of the commonly used RFM model. RFM stands for Recency, Frequency, and Monetary Value, and it is often used in marketing to identify the best customers. In our case, the customers are the blood donors themselves.

Here is a simplified explanation of each column in the dataset:

- Recency (R): This column represents the number of months since the donor's last blood donation.
- Frequency (F): It indicates the total number of times the donor has donated blood in the past.
- Monetary (M): This column represents the total amount of blood donated by the individual, measured in cubic centimeters (c.c.).
- Time (T): It indicates the number of months since the donor made their first blood donation.
- Target: This is a binary variable that indicates whether the donor donated blood in March 2007. A value of 1 represents donating blood, while 0 represents not donating blood.

By analyzing this dataset, we aim to gain insights and develop a predictive model to determine the likelihood of a donor giving blood in the future visit of the blood donation vehicle. The RFMTC structure allows us to consider factors such as recency, frequency, monetary value, and the time since the first donation, which can be helpful in predicting future donation behavior.

3.2 Statistical Insights of Dataset

By conducting statistical analysis on our dataset, we have discovered some valuable insights. Here are the key findings:

1. All columns in the dataset contain numerical data. This means that the information in the dataset is represented using numbers rather than text or categorical variables.
2. When examining the target column, which indicates whether a donor donated blood in March 2007, we observed the following distribution
    a. Approximately 76.2% of the records have a value of 0, indicating no blood donation,
    b. while around 23.8% of the records have a value of 1, indicating a blood donation. This shows that there is an imbalance in the distribution of the target classes.

```
In [9]: # Print target incidence proportions, rounding output to 3 decimal places

        target_proportions = round(transfusion['target'].value_counts(normalize=True), 3)

        # Print target incidence proportions
        target_proportions

Out[9]: 0    0.762
        1    0.238
        Name: target, dtype: float64
```

3. Due to the uneven distribution of the target classes, we performed a dataset split for training and testing purposes. The parameters used for the split were as follows:
    a. The test dataset was assigned a size of 25% of the total data. A random state value of 42 was set to ensure reproducibility of the results.

b. Additionally, the stratify parameter was applied based on the target column to maintain the proportional representation of the target classes in both the training and testing datasets.

These statistical insights provide us with a better understanding of the dataset's characteristics and help guide our further analysis and modeling decisions.

3.3 Model Selection and Development

To simplify the process of selecting the most suitable algorithm and developing the entire pipeline, we utilize an automated machine learning (AutoML) tool called TPOT. TPOT, which stands for "The Tree-Based Pipeline Optimization Tool," is a Python library that optimizes machine learning pipelines using genetic programming.

TPOT automates the time-consuming part of machine learning by intelligently exploring numerous possible pipelines and selecting the best one for our dataset. It is built on top of the popular scikit-learn library, so the code it generates should be familiar to users.

After evaluating various pipelines and algorithms, TPOT determined that "Logistic Regression" is the best model for our dataset. No additional preprocessing steps were required. The model achieved an AUC score of 0.7850, indicating its effectiveness in predicting blood donation behavior.

Based on these findings, we have chosen to use the Logistic Regression algorithm for the development of our predictive model.

3.4 Model Training & Evaluation

In linear regression models, one of the assumptions is that the relationship between the data and features follows a linear pattern or can be measured using a linear distance metric. However, if a feature in our dataset has a significantly higher variance compared to other features, it can affect the model's ability to learn from the remaining features.

To address this issue, we need to normalize the data, which is a technique used to correct for high variance. In our case, the 'Monetary' column, representing the total

blood donated in cubic centimeters (c.c.), had a significantly higher variance compared to other columns. If left unaccounted for, this feature might be given more weight by the model and be seen as more important than other features. To address this, we applied log normalization.

The logistic regression model was trained using the following parameters:

1. **Solver: liblinear**
2. **Random State: 42**

To evaluate the model's performance, we calculated the AUC (Area Under the Curve) score, which is a common metric used to assess the quality of a classification model. In our case, the logistic regression model achieved an AUC score of **0.7891**, indicating its effectiveness in predicting blood donation behavior.

In [16]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score

# Instantiate LogisticRegression
logreg = LogisticRegression(
    solver='liblinear',
    random_state=42
)

# Train the model
logreg.fit(X_train_normed, y_train)

# AUC score for logreg model
logreg_auc_score = roc_auc_score(y_test, logreg.predict_proba(X_test_normed)[:, 1])
print(f'\nAUC score: {logreg_auc_score:.4f}')
```

```
AUC score: 0.7891
```

You can find the code for this project in the following Google Drive link:

https://drive.google.com/drive/folders/1xuRXMOb2eayn3Y5pf1ZFdU40_9KHHapF?usp=drive_link

CONCLUSION & FUTURE SCOPE

- The demand for blood varies throughout the year, with donations slowing down during busy holiday seasons.
- Forecasting the future supply of blood is crucial to taking proactive actions and saving lives.
- We used TPOT and achieved an AUC score of 0.7850, which is better than randomly choosing '0' all the time (which would have a success rate of 76% based on target incidence).
- Log normalization of the training data further improved the AUC score by 0.5%.
- Logistic regression models offer interpretability, allowing us to analyze how much of the variance in the target variable can be explained by other variables in the dataset.
- Pre-donation information and counseling play a vital role in the process of donor selection, ensuring that individuals meet the necessary criteria based on their medical history.
- Providing donors with information about the blood donation process, including selection criteria, deferral options, blood screening, and counseling, allows individuals unsuitable for donation to self-defer without undergoing the 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/