Final Report: Predicting Blood Donation Using Machine Learning

ACKNOWLEDGMENTS

The virtual traineeship opportunity that I had with MedTourEasy was a great chance for learning and understanding the intricacies of machine learning in data analytics, particularly in blood donation forecasting. It was also a valuable experience for both personal and professional development. I am very grateful for the opportunity to interact with professionals who guided me throughout this project and helped make it a great learning experience.

Firstly, I express my deepest gratitude and special thanks to the Training & Development Team of MedTourEasy for providing me with the opportunity to carry out this traineeship at their esteemed organization. I also extend my thanks to my colleagues for helping me understand the details of data analytics and machine learning. Their guidance enabled me to successfully complete this project with a strong foundation in data-driven decision-making and predictive modeling.The traineeship opportunity that I had with this project was a great chance for learning and understanding the intricacies of machine learning in data analytics, particularly in blood donation forecasting. It was also a valuable experience for both personal and professional development. I am very grateful for the opportunity to interact with professionals who guided me throughout this project and helped make it a great learning experience.

Firstly, I express my deepest gratitude and special thanks to my mentor and the support team who provided me with an opportunity to carry out this project. I also extend my thanks to my colleagues for helping me understand the details of data analytics and machine learning. Their guidance enabled me to successfully complete this project with a strong foundation in data-driven decision-making and predictive modeling.I would like to express my sincere gratitude to my mentor, colleagues, and all those who guided me throughout this project. Their continuous support and encouragement helped me gain valuable insights into data analysis and machine learning. I appreciate the opportunity to work on this project, which has been an enriching learning experience.

TABLE OF CONTENTS

ABSTRACT

Blood transfusion is a critical component of healthcare, ensuring that patients in need receive timely and sufficient blood supply. This project aims to predict whether a donor will donate blood again in the future. Using machine learning techniques, we analyze historical donation data to improve forecasting and optimize blood collection efforts. We employed logistic regression and an automated machine learning approach (TPOT) to identify the best predictive model. The results indicate that machine learning can significantly enhance decision-making in blood donation management.

1. INTRODUCTION

1.1 About the Project

The goal of this project is to build a machine learning model that predicts whether a blood donor will donate again within a given time frame. The dataset, obtained from the Machine Learning Repository, consists of a random sample of 748 donors from a mobile blood donation vehicle in Taiwan. The dataset follows the RFMTC (Recency, Frequency, Monetary, Time, and Compactness) model used in marketing analytics.

1.2 Objectives and Deliverables

- Analyze donor data and extract meaningful insights.
- Preprocess and clean the dataset for accurate predictions.
- Implement machine learning models to predict future blood donations.
- Compare the performance of different models.
- Provide recommendations for blood banks based on model predictions.

2. METHODOLOGY

2.1 Flow of the Project

The project followed these key steps:

1. **Understanding the Dataset** – Analyzing data structure, missing values, and distributions.
2. **Data Preprocessing** – Cleaning and normalizing data to enhance model accuracy.
3. **Feature Engineering** – Identifying and transforming key variables.
4. **Model Selection** – Training different machine learning models.
5. **Evaluation and Optimization** – Measuring model performance using metrics like AUC.
6. **Final Model Deployment** – Selecting the best-performing model.

2.2 Tools and Technologies Used

- **Programming Language**: Python
- **Libraries**: Pandas, NumPy, Scikit-learn, TPOT
- **Machine Learning Models**: Logistic Regression, Automated TPOT Classifier

3. IMPLEMENTATION

3.1 Data Collection and Importing

- The dataset was loaded using Pandas and inspected using head(), info(), and describe().
- The target variable (whether a donor donated in March 2007) was renamed for better readability.

3.2 Data Preprocessing and Cleaning

- Checked for missing values (none found).
- Analyzed class distribution to handle imbalances.
- Log-normalized high-variance features for better model performance.

3.3 Exploratory Data Analysis

- Visualized donor trends to understand behavior patterns.
- Identified correlations between features using heatmaps.

3.4 Model Selection and Training

- **TPOT Classifier**: Used genetic algorithms to automate model selection.
- **Logistic Regression**: Implemented as a baseline model.
- The dataset was split into **75% training** and **25% testing** using train_test_split().

3.5 Model Evaluation

- **TPOT AUC Score**: 0.7850
- **Logistic Regression AUC Score**: Slightly lower after normalization.
- **Comparison of models:** TPOT performed better in selecting an optimized pipeline.

4. RESULTS AND OBSERVATIONS

4.1 Model Performance Summary

Model	AUC Score
TPOT Classifier	0.7850
Logistic Regression	Slightly lower

- TPOT provided an automated and optimized machine learning pipeline.
- Logistic regression, while interpretable, had slightly lower accuracy.
- Normalization improved model performance.

5. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This project demonstrated the potential of machine learning in predicting future blood donations. The TPOT classifier provided the best results, showcasing the power of automated model selection. The findings can help blood banks optimize their collection drives and ensure sufficient supply.

5.2 Future Scope

- **Implementing additional models** such as Random Forest and XGBoost.
- **Expanding the dataset** to improve model generalization.
- **Exploring time-series forecasting** to analyze donation trends.
- **Integrating real-time data** from donation centers for dynamic predictions.

6. REFERENCES

- Machine Learning Repository Dataset: MedTourEasy Learning Platform
- TPOT Documentation: [https://epistasislab.github.io/tpot/]
- Scikit-learn Documentation: [https://scikit-learn.org/]

Prepared by: Lakshya Rastogi
Date: 28/02/2025