UNIVERSITY OF LONDON

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

MACHINE LEARNING REPORT

Realized by: Ibtihel Frini

03 AVRIL 2025

LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE
Student ID: 220565039

Executive Summary

In order to better understand the behavior of donors, this paper offers a thorough examination of blood donation trends using machine learning approaches. Three fundamental machine learning tasks were used to evaluate the dataset, which was acquired from the **UCI Blood Transfusion Service Center**: regression, classification, and clustering (unsupervised learning).

• Based on their past donations, Task 1 (Clustering) distinguished three different donor groups: frequent/long-term donors, infrequent/new donors, and moderate donors. Blood donation facilities can use this segmentation to better target their engagement methods for various donor types.

• Task 2 (Regression) , with Random Forest Regression emerging as the top model, showed that donation frequency can be predicted with moderate accuracy, highlighting the importance of donor engagement duration and recency.

• To determine if a donor will make another donation, Task 3 (Classification) employed logistic regression, decision trees, and random forests. Logistic Regression (Balanced) achieved the highest recall, effectively identifying potential donors. However, it faced difficulties predicting donors who were likely to donate again, highlighting the challenges in classifying repeat donors.

Blood donation facilities can benefit from this data by creating focused interventions, increasing donor retention, and making more accurate predictions about future donations.

Table of Contents

Table of figures

1. Introduction

Blood donation facilities worldwide struggle to keep a consistent supply of blood, particularly because donor retention rates might change over time. In order to engage, retain, and forecast future donation trends, it is imperative to comprehend the behavior of donors. This study uses a variety of machine learning approaches to examine donor behavior using a dataset from the UCI Blood Transfusion Service Center. Three tasks comprise the analysis:

1. Clustering to identify homogeneous population groups.
2. Regression to predict the frequency of donations.
3. Classification to predict whether a donor will donate again.

Each task provides valuable insights that can be utilized to improve blood collection tactics and more effectively target donors.

2. Dataset Overview

The dataset consists of four main features:

- **Recency**: Time (in months) since the last donation.
- **Frequency**: Total number of donations.
- **Monetary**: Cumulative amount of blood donated.
- **Time**: Time (in months) since the first donation.
- **Class**: Whether the donor donated blood in March 2007 (1 = Yes, 0 = No).

The dataset is suitable for clustering (to group similar donors), regression (to predict donation frequency), and classification (to predict future donations).

3. Task 1: Clustering Analysis

Before proceeding with the clustering and subsequent analysis, the dataset was standardized. This step ensured that all features were on the same scale, eliminating any bias that could arise from differing units or ranges of values.

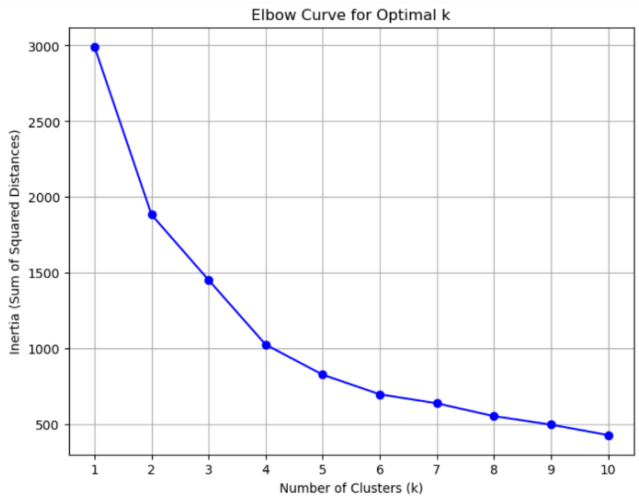


Figure 1: Elbow Curve

The elbow curve in figure 1 suggests that the optimal number of clusters is **3**, as beyond this point, the reduction in within-cluster sum of squares (WCSS) becomes less significant. This supports our segmentation of donors into **Frequent and Long-Term Donors, Infrequent/New Donors, and Moderate Donors**, allowing for more targeted engagement strategies.

3.1 Cluster Centroids (Mean of Numeric Columns)

	Recency	Frequency	Monetary	Time
Cluster 0	7.964	10.12	2530.97	59.68
Cluster 1	10.27	2.89	722.22	22.08
Cluster 2	4.67	39.44	9861.11	91.89

3.2 Interpretation of Clusters

- **Cluster 0 (Frequent and Long-Term Donors):** Donors in this cluster have a high donation frequency (10.12) and high cumulative donation amount (2530.97) over an extended time period (59.68 months). Their recency (7.964 months) is moderate, indicating they have donated relatively recently. These donors are likely committed and regular contributors.

- **Cluster 1 (Infrequent/New Donors):** This group consists of donors who have a low frequency (2.89 donations) and lower monetary contributions (722.22). Their recency (10.27 months) is the highest among all clusters, meaning they haven't donated recently. They also have the shortest donation history (22.08 months), suggesting they are either new donors or donate sporadically.

- **Cluster 2 (Highly Active Donors):** These donors stand out due to their very high frequency (39.44 donations) and the highest cumulative donation amount (9861.11). Their recency (4.67 months) is the lowest, meaning they donate frequently and consistently. They also have the longest donation history (91.89 months), indicating a strong, sustained engagement with the donation program.

3.3 Research Question

• Can we determine distinct donor segments using **KMeans clustering and PCA**, as indicated by the **elbow curve**, to develop personalized engagement strategies for each group?

3.4 KMeans Clustering of Donors (PCA)

The correlation circle plot helps explain how the PCA components relate to the original features in the dataset:

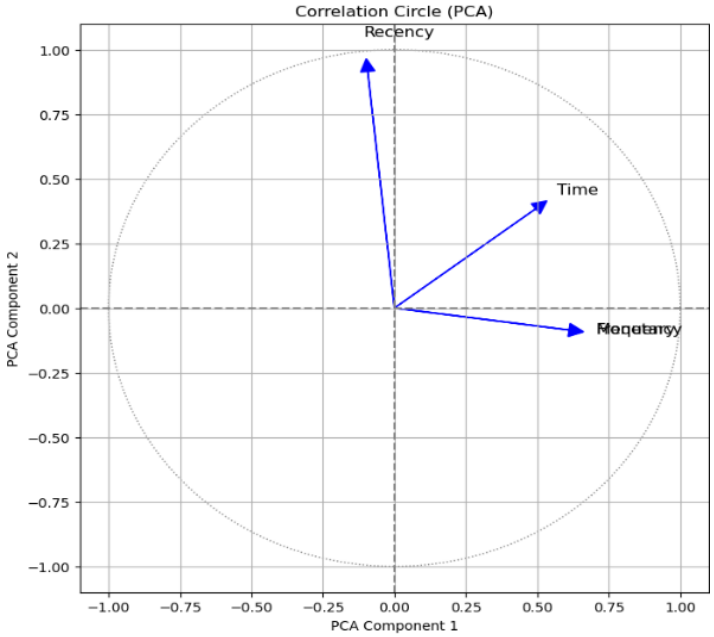


Figure 2: Correlation Circle

- **PCA Component 1** is strongly correlated with **Monetary** and **Frequency**, as indicated by the arrows pointing in the same direction. This means **Component 1 captures how often and how much a donor has contributed**.

- **PCA Component 2** is strongly associated with **Recency**, meaning **Component 2 represents how recently the donor has given blood**.

- **Time is positioned between PCA Component 1 and PCA Component 2**, indicating that it has a mixed effect on both principal components but leans more towards donation frequency.

Since PCA Component 1 is associated with **donation history (frequency and monetary amount)** and PCA Component 2 is linked to **recency**, the clusters in the scatter plot can be understood as follows:

- **Cluster 0 (purple, moderately active donors) is spread across the middle** as they have a mix of moderate recency and frequency.
- **Cluster 1 (green, infrequent donors) is positioned on the left side** of the scatter plot because they have low PCA Component 1 values (low frequency and monetary contribution) but higher recency values.
- **Cluster 2 (yellow, highly frequent donors) is positioned far on the right** because they have the highest PCA Component 1 values (frequent donations) and lower recency.

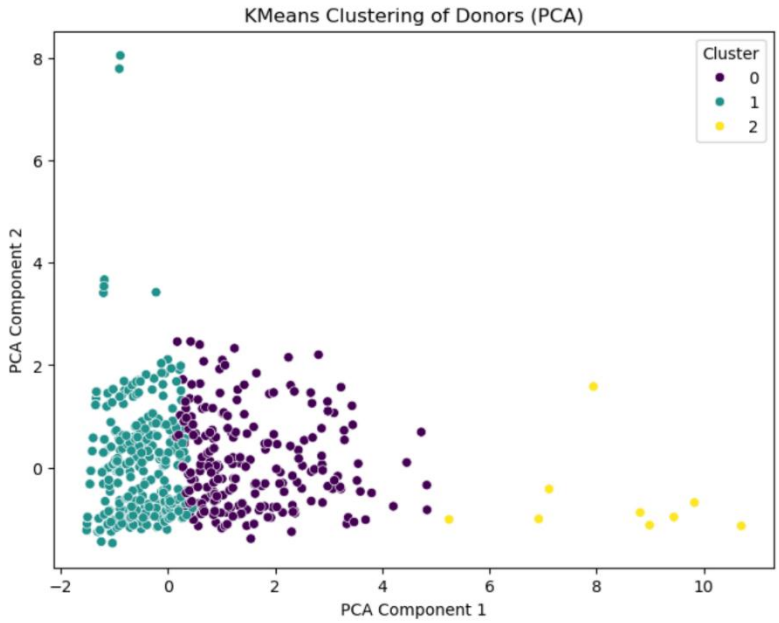


Figure 3: KMeans Clustering

The scatter plot clearly distinguishes three donor groups based on their donation behavior after applying KMeans clustering with PCA:

• **Cluster 0 (Frequent and Long-Term Donors):** These donors show a moderate recency of 7.96 months and have donated around 10 times with a cumulative donation of 2530.97. They are relatively consistent in their donations but not the most frequent. They likely represent engaged donors who can be nurtured for continued contributions.

• **Cluster 1 (Infrequent/New Donors):** This cluster contains donors who have donated fewer times (2.88 donations on average) and have a long recency of 10.27 months since their last donation. Their donation amounts are relatively low, indicating that they might be new donors or those who donate infrequently. Targeting these donors for re-engagement campaigns could encourage them to donate more frequently.

• **Cluster 2 (Highly Active Donors):** Donors in this group have the highest frequency (39.44 donations) and cumulative donation amount (9861.11), with a recency of just 4.67 months. They are the most committed donors, contributing consistently over time. Maintaining their involvement should be a priority, as their behavior suggests long-term loyalty.

These groups highlight the different engagement levels among donors, enabling blood donation facilities to create targeted strategies: re-engagement efforts for Cluster 1, and retention programs for Clusters 0 and 2.

4. Task 2: Regression Analysis

In this section, we aim to predict the **donation frequency** of donors, which is essential for understanding long-term donor engagement and optimizing fundraising strategies. To achieve this, we use the following features as predictors:

- **Recency:** The time since the donor's last donation.
- **Total Donations:** The total number of times a donor has contributed.
- **Time Since First Donation:** The duration between the donor's first recorded donation and the present.
- **Monetary Contributions:** The total amount donated by the donor.

These features help assess donor behavior and determine the likelihood of continued contributions.

4.1 Model Performance

Model	R-squared (R^2)	Root Mean Squared Error (RMSE)	Root Mean Absolute Error (MAE)
Linear Regression	0.426545	3.512624	1.500209
Decision Tree Regression	0.311310	3.849407	1.513367
Random Forest Regression	0.507630	3.254828	1.427053
KNN Regression	0.416954	3.541876	1.477235
AdaBoost Regression	0.395102	3.607638	1.611441

Based on the model performance metrics, **Random Forest Regression** emerges as the best-performing model, with the highest **R^2 score** of 0.51, indicating it explains 51% of the variance in the data. It also has the lowest **Root Mean Absolute Error (RMAE)** of 1.43, suggesting that its predictions are closest to the true values on average.

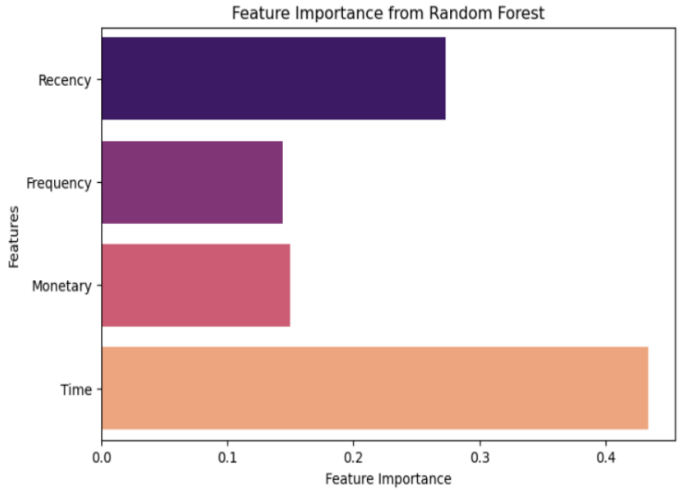
In comparison, **Decision Tree Regression** exhibits the lowest **R^2 score** of 0.31, indicating it explains only 31% of the variance in the data, which suggests weaker model fit. It also has a relatively high **RMAE** of 1.51, pointing to less accurate predictions compared to the other models.

Overall, the **Random Forest Regression** model outperforms the others both in terms of **fit** and **prediction accuracy**, with a balance of relatively high **R^2** and lower **RMSE** and **RMAE**.

4.2 Research Question

• How do different donor characteristics (such as recency, total donations, and time since the first donation) influence the frequency of future donations?

4.3 Feature Importance (Random Forest Analysis)



The feature importance graph provides insights into how different donor characteristics—**recency, total donations (monetary), and time since the first donation**—affect the frequency of future donations.

Figure 4: Random Forest Feature Importance

According to the Random Forest model, **time since the first donation** (≈ 45%) is the most influential factor, suggesting that donors who have been engaged for a longer period are more likely to donate frequently. **Recency** (≈ 30%) also plays a crucial role, indicating that donors who have given recently are more likely to continue donating. Meanwhile, **monetary contributions** and **past donation frequency** have smaller but still relevant impacts. This suggests that while higher donation amounts and frequent past donations contribute to predicting future giving behavior, they are less critical compared to engagement duration and recency.

These findings highlight that donor retention strategies should focus on maintaining long-term relationships and ensuring timely follow-ups after donations. Organizations can optimize outreach campaigns by targeting donors who have been engaged for a long time and those who have recently donated, ultimately enhancing fundraising success.

4.4 Limitations and Insights

The feature importance analysis demonstrates that donor characteristics—particularly **time since the first donation** and **recency**—strongly influence the frequency of future donations. The findings align with prior research emphasizing the role of engagement duration and recent interactions in predicting donor behavior **(Fader & Hardie, 2009).** However, certain limitations must be considered.

While the Random Forest model provides meaningful insights, its feature importance rankings do not capture potential interactions between variables. Additionally, the model's reliance on historical data means that external factors, such as economic conditions or shifts in donor motivations, are not accounted for **(Sargeant & Woodliffe, 2007).**

To enhance generalizability, the model should be validated on new datasets and tested across different donor segments. Future research could explore alternative modeling techniques, such as time-series forecasting or causal inference methods, to better understand the underlying drivers of donation frequency.

5. Task 3: Classification Analysis

In this section, we develop models to predict whether a donor will **donate again** based on past donation behavior. Accurately forecasting repeat donors is crucial for donor retention strategies and effective outreach campaigns. The following features are used for prediction:

- **Past Donation Frequency:** The number of times a donor has donated in the past.
- **Time Since Last Donation (Recency):** The gap between the most recent donation and the present.
- **Total Donations:** The total number of donations made by the individual.
- **Engagement Duration:** The time elapsed since the donor's first contribution.

By analyzing these variables, we aim to identify patterns in donation behavior and improve future fundraising efforts.

5.1 Model Performance

The table below presents the performance metrics for several classification models, including accuracy, F1 score, precision, recall, and AUC-ROC. These metrics provide insight into the models' ability to correctly classify instances and their overall effectiveness in distinguishing between classes.

Model	Accuracy	F1 Score	Precision	Recall	AUC-ROC
Logistic Regression (Balanced)	0.657778	0.549708	0.423423	0.783333	0.761061
Logistic Regression	0.737778	0.144928	0.555556	0.083333	0.757929
Decision Tree	0.666667	0.311927	0.346939	0.283333	0.564747
Random Forest	0.733333	0.347826	0.500000	0.266667	0.677525
KNN	0.702222	0.336634	0.414634	0.283333	0.666010
SVC	0.737778	0.032787	1.000000	0.016667	0.626515
QDA	0.488889	0.449761	0.315436	0.783333	0.600808
AdaBoost	0.760000	0.341463	0.636364	0.735505	0.735505

In the context of blood donation prediction, recall is particularly important since it ensures that potential donors who are likely to donate are correctly identified.

Logistic Regression (Balanced) achieves the highest recall (0.7833) among the models with a decent precision (0.4234), meaning it effectively captures actual donors while keeping false positives at a manageable level.

While other models, such as SVC, show high precision, their low recall makes them impractical, as they fail to identify most potential donors. AdaBoost, despite having the highest accuracy (76%), sacrifices recall and balance.

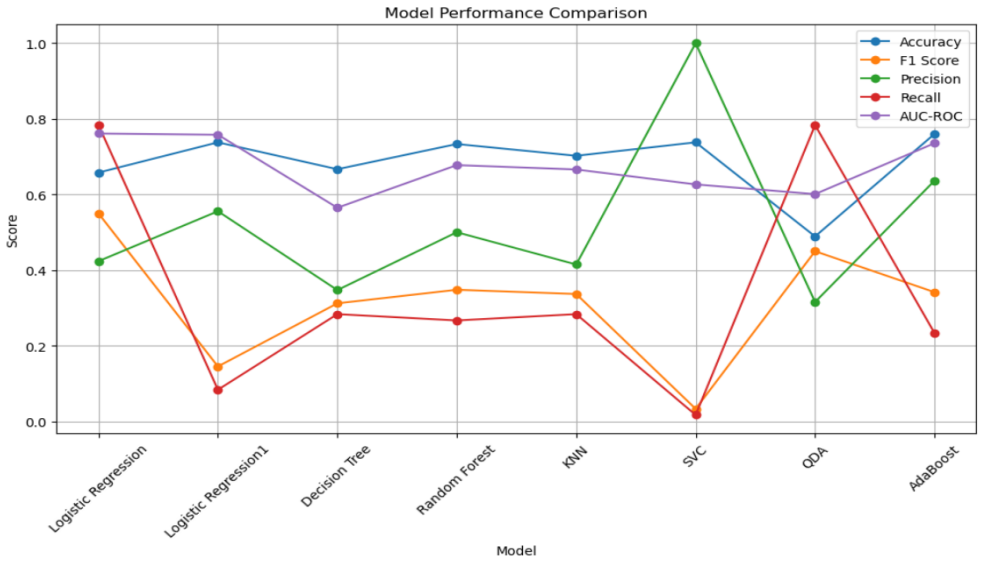


Figure 5: Model Performances

Since the goal is to maximize donor identification while maintaining reasonable accuracy, **Logistic Regression (Balanced)** is the best choice, ensuring that fewer potential donors are overlooked while keeping misclassifications under control.

5.2 Logistic Regression Classification Report:

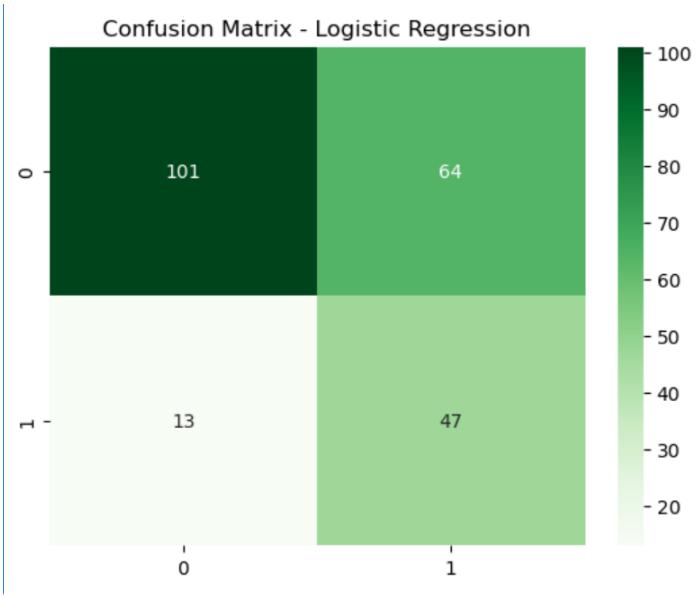
	Precision	**Recall**	**F1-Score**	**Suuport**
Accuracy			0.66	225
Class 0	0.89	0.61	0.72	165
Class 1	0.42	0.78	0.55	60

The Logistic Regression model shows high precision for Class 0 (0.89) but lower recall (0.61), indicating it predicts non-donors accurately but misses some. For Class 1, recall is higher (0.78), but precision is low (0.42), meaning it identifies many donors but also misclassifies non-donors as donors. Overall accuracy is 66%, but the model struggles with predicting donors.

5.3 Research Question

• Is it possible to forecast a donor's propensity to donate again based on past donations?

5.4 Confusion Matrix (Logistic Regression)



According to the confusion matrix, historical contribution trends can be used to forecast a donor's likelihood of making another donation. The model's ability to predict outcomes is illustrated by the Confusion Matrix for Logistic Regression (Balanced).

Figure 6: Confusion Matrix

For class 0 (donors who are predicted **not to donate again**), there were 101 correct predictions and 13 incorrect predictions. For class 1 (donors who are predicted **to donate again**), 47 were correctly predicted, while 64 were incorrectly predicted.

This indicates that the model is more successful at predicting non-donors (class 0), with a relatively high accuracy of 101 out of 114. However, it struggles to predict those who are likely to donate again (class 1), with a higher number of false negatives (64), meaning the model is underestimating the likelihood of donors returning.

Overall, the algorithm seems to perform better in identifying donors who will stop donating but struggles with forecasting repeat donors.

5.5 Suggestions for Improvement

Because of the class imbalance, the model did well in predicting non-donors (Class 0) but poorly in predicting donors (Class 1). As noted by **King and Zeng (2001)**, class imbalance is a common challenge when predicting rare events, where one class is significantly underrepresented. This imbalance can lead to the model's bias toward predicting the majority class (non-donors), while underperforming on the minority class (donors). According to **Chawla et al. (2002)**, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) can be used to address this issue by generating synthetic samples for the minority class, improving the model's ability to accurately predict donors.

6. Conclusion and Perspectives

This analysis emphasizes the significance of using machine learning in forecasting donor behavior:

- **Through clustering**, different donor segments were identified, providing valuable insights for tailored engagement strategies that could optimize outreach.
- **Regression**, with Random Forest as the best-performing model, demonstrated that donation frequency can be highly predictable based on past data, offering a strong foundation for forecasting donor behavior.
- **Classification** showed that Logistic Regression (balanced) can reliably predict donors; however, further work is required to predict non-donors, especially due to the class imbalance issue.

By integrating these findings, blood donation facilities can enhance donor retention, improve targeted engagement efforts, and ensure a consistent and reliable donation flow.

7. References

- **UCI Machine Learning Repository**. Blood Transfusion Service Center Data Set. Available at: https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center
- **Fader, P. S., & Hardie, B. G.** (2009). Probability Models for Customer-Base Analysis. Journal of Interactive Marketing, 23(1), 61-69.
- **Sargeant, A., & Woodliffe, L.** (2007). Gift giving: An interdisciplinary review. International Journal of Nonprofit and Voluntary Sector Marketing, 12(4), 275-307.
- King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2), 137-163.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.