Predicting Blood Donor Retention with Light GBM: A High-Performance Gradient Boosting Framework

Nahashon Kiarie
School of Computing and Informatics. Meru University of Science and Technology, Meru, Kenya

Mary Mwadulo
School of Computing and Informatics. Meru University of Science and Technology, Meru, Kenya

Amos Chege Kirongo
School of Computing and Informatics. Meru University of Science and Technology, Meru, Kenya

ABSTRACT

Blood donation is critical for ensuring a stable and reliable supply of blood, yet blood donor retention remains a complex and persistent challenge. Previous attempts to develop predictive models for blood donor retention have often yielded relatively low accuracy and fail to address the class imbalance challenge that come with blood donation data, limiting their practical application in addressing this challenge. This study investigates the use of the Light Gradient Boosting Machine (LightGBM) as a high-performance gradient boosting framework for predicting blood donor retention. LightGBM employs a leaf-wise growth strategy, which significantly improves accuracy by minimizing loss at each iteration. It also supports histogram-based learning, reducing memory consumption and speeding up computation, making it suitable for the blood donation prediction. The study utilized data obtained from Kenya blood banks, consisting of 5000 records and nine features, to develop and evaluate the model. The LightGBM model achieved an accuracy of 98.3% and an F1 score of 97.8 which was higher as compared to the existing models. The results demonstrate that LightGBM is an effective and computationally efficient tool for predicting blood donor retention. Its ability to handle large, imbalanced datasets and complex patterns makes it well-suited for real-world applications in predictive analytics. This study provides blood agencies with a more reliable model for accurately predicting blood donor retention, reducing recruitment costs, and enabling targeted retention strategies to ensure a steady blood supply.

General Terms

Model, Machine learning, Algorithms. Predictive Models.

Keywords

Blood donation, Light GBM, Gradient Boosting, Blood donor retention.

1. INTRODUCTION

Blood is an essential component of the human body responsible for transportation of nutrients, oxygen, hormones and other elements necessary for proper functioning of the body. A healthy blood supply is essential for effective body performance. Donor retention is the process of retaining blood donors to give blood regularly[1]. However, blood donation centers face challenges due to declining blood donor retention rates, hence reducing their ability to provide sufficient blood[1]. One of the major factors that contribute to blood shortages is the high rate of blood donor attrition, with many first-time blood donors failing to return for subsequent blood donations. Recruiting new donors is often expensive and time-consuming [2]. Enhancing donor retention is very crucial in ensuring a stable supply of blood, reducing costs associated with recruitment, and improving the overall healthcare outcomes.

Machine learning has become a transformative force in healthcare enabling transformations and unlocking limitless possibilities [3]. One of the key benefits of machine learning in healthcare is its ability to analyse huge amounts of complex data, including electronic health records, medical images, genetic information, real-time patient monitoring data among other medical data. Machine learning models are used to analyze historical patient data, they can be able to predict the likelihood of disease outbreaks, patient admissions and readmissions, and any adverse drug reactions to the patients[4]. By utilizing the power of technology, the machine learning predictive models can help to identify key factors that influence blood donor retention and hence enable personalized donor retention strategies tailored towards individual donor's needs and preferences[5].

While machine learning models have been applied to predict blood donor retention, existing models often suffer from significantly low predictive accuracy[6][7]. Moreover, these models often struggle to address critical challenges that come with blood donation datasets, such as class imbalance, which frequently results in overfitting and reduced generalization performance[8].

The objective of this study was to develop a machine learning model for predicting blood donor retention based on lightGBM as a high-performance gradient boosting framework.

This study is justified by the urgent need to address the issue of blood donor retention and the need to provide a more accurate and intelligent-based solution.

2. REVIEW OF RELATED LITERATURE

Blood donor retention is the ability of blood centers to keep donors active and prevent them from lapses in their blood donation. It is a more cost-effective way of retaining active blood donors as opposed to recruiting new donors, this strategy is essential for ensuring the continuity and safety of the blood supply[1]. Blood has a short life span and cannot be manufactured in laboratories, its demand is very high and therefore its supply should always remain constant. Emergency situations such as accidents, medical operations and diseases necessitate regular blood transfusion[9]. The demand for blood and blood products is constantly increasing due to population growth, advancements in medical procedures, and rising incidence of diseases such as cancer and chronic conditions that require regular transfusions[10]. However, this increasing demand is not being met adequately, resulting in blood shortages and their subsequent impact on healthcare systems worldwide[11]. Machine learning functions by learning data

and generating prediction rules by recognizing patterns in the data, as opposed to following a predefined and hard-coded algorithm. The recursive nature of machine learning allows it to adapt and evolve in response to new data changes[12]. Machine learning algorithms have been used in various studies to predict blood donations and blood donor retention.

In their study on forecasting blood donor response using predictive modelling approach[6] used predictive modeling approach to predict whether a particular donor will donate blood within coming months. The study used existing dataset obtained from the open database of Blood Transfusion Service Centre in Taiwan. The study compared various classification algorithms such as K-nearest Support vector machines, Neighbours (KNN), Decision tree, Gaussian Naive Bayes, and logistic regression. The results show that decision tree produced the best accuracy at 0.60.

The study[13] classified eligibility of blood donors using decision trees and Naive Bayes classifiers. The study employed a data set of 500 blood donors, obtained from a humanitarian organization in Indonesian. The decision tree classifier achieved an accuracy of 78.5%, while Naive Bayes classifier achieved an accuracy of 81.5%.

The study on analytics framework for blood donor classification in 2021 classified students from an Indian state university as potential blood donors or non-donors using data visualization techniques [7]. KNN classifier produced the best results with an Accuracy of 0.7027, Precision of 0.7209, Sensitivity value 0.7949, F1-score equal to 0.7561 and Specificity value of 0.5789.

While predicting the return rate in young blood donors the study [5] by Cloutier, in 2021 extracted data from a blood donation management information system managed by Héma-Québec. The dataset analyzed included 81 986 donors aged 18–24 at the time of their most recent donation. The data contained 11 main attributes. The study employed Random forest and mean decrease accuracy (MDA) method to measure the features impact on the accuracy of the model and cross validation was used to validate the model. The random forest model accurately predicted over 91% of donation frequencies, with an overall average error rate of 8.16% and specific error rates of 4.6% and 12.3% for the 'unreturned donor and returned donor groups respectively

The study [14] done in 2022 aimed at building a forecasting system for donation of blood using SVM Model, obtained data from a Blood Transfusion Service Center in Hsin-Chu City in Taiwan. The dataset included 748 donors with five main variables. Support Vector Classifier obtained the highest accuracy of 78.4%.

3. METHODOLOGY

3.1. Experimental Setup

The study adopted an experimental research design to build, test and validate the model. The experimental set up followed the following steps: The dataset was extracted, Data preprocessing was done to clean the data, check missing values and detect outliers, feature selection was done to select the best features for model training, data was split into training and testing. The Light GBM model was trained using cross validation, the model hyperparameter tuning was done and finally the performance evaluation for the model was done to produce the final model which was compared to the existing models.

Figure 1 below show the experimental set up stages and flow.

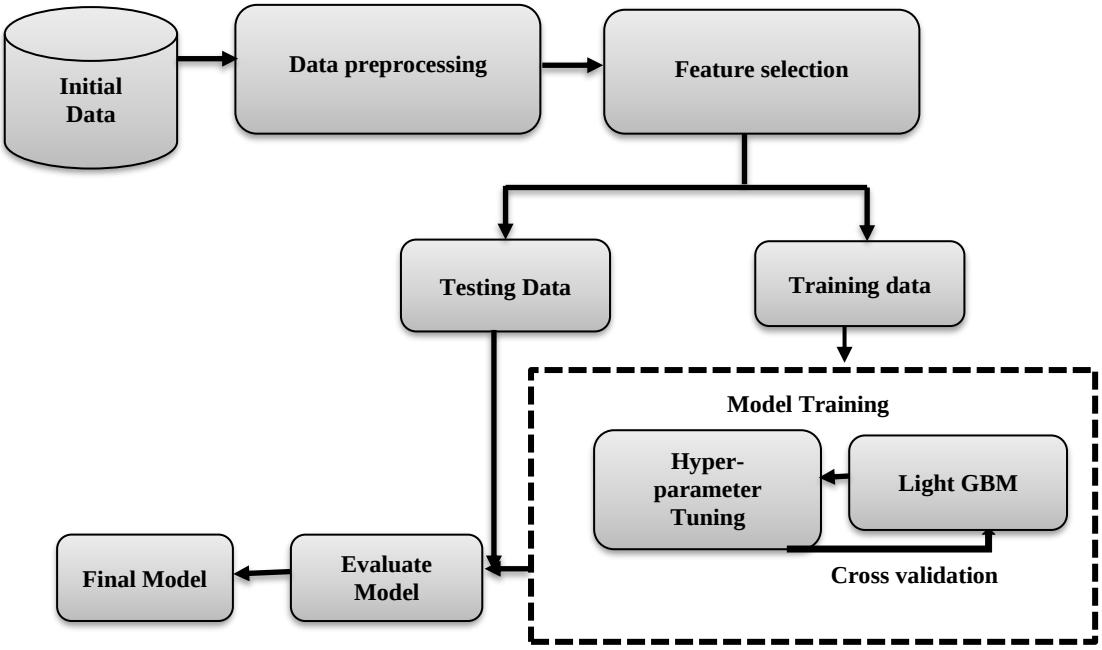


Figure 1. Experimental Set Up

3.2.The Dataset

Data used in this study was obtained from Kenya blood bank management system it consists of blood donors registered in the system from the year 2022. The data consists of 5000 records with nine (9) features. The dataset includes both numerical and categorical variables. categorical variables were encoded before being fed into the classifiers. The Data preparation involved cleaning, converting, and organizing the

raw data into a format that can be analyzed and modeled. Table 1 below shows the description of the variables in the dataset

Table 1. Data Description

Attribute	Description	Values
Gender	Gender of the door	Male, Female
Age	Age of the blood donor	Number 17-65
Education Level	The highest education level achieved.	None, Primary, Secondary., Tertiary.
Blood group.	The blood group of the blood donor.	A+, A-, B+, B-, AB+, AB-, O+, and O
Months since last donation	Total number of months since last donation Number	0 and above
No of Previous donations	Total number of donations made by the donor including the current donation Number.	0 and above
Months since First Donation	Total number of months since the donor made the first donation Number	0 and above
Total Volume donated	Total volume of blood that the donor has donated since they started donating blood. Number	0 and above
Donated Blood in 2024	Binary variable indicating whether a donor has donated blood in 2024	0 or 1

3.3. Model Training

A systematic training process was followed to rigorously train the model. The dataset was initially divided into 80% training data and 20% testing data. This gives the model enough data for training allowing it to learn the underlying trends and patterns while reserving a significant portion for testing[15].

The light gradient boosting algorithm was imported and trained using the training data and based on its default parameters. K fold cross validation was utilized with k=5. This means that the training dataset was split into 5 equal-sized folds, and the model was trained and evaluated 5 times, each time using a different fold as the validation set and the remaining folds as the training set. This process allows for a more reliable estimation of the model's performance compared to a single train-test split[16].

The model was developed using python packages. Jupiter notebook was used as the platform for coding the python program. The python libraries utilized include: Numpy, Pandas, scikit-learn, matplolib and seaborn.

4. RESULTS AND DISCUSSIONS

4.1. Exploratory Data Analysis

Statistical summary of the data of the numerical values was done using python. The structural analysis of the data shows the mean, standard deviation, minimum, maximum, 25th percentile and 75th percentile for each of the numerical variables. The mean shows the average of the values while the standard deviation measures how the numerical values were spread. Figure 2. below shows the statistical summary of the numerical variables

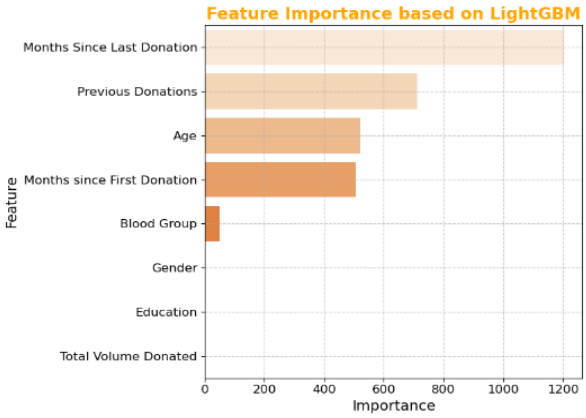
Figure 2. Statistical Summary of the Numerical Variables

Statistic	Age	Months Since Last Donation	Previous Donations	Months Since First Donation	Total Volume Donated	Donated Blood in 2024
Count	5000	5000	5000	5000	5000	5000
Mean	23.95	13.40	2.37	20.40	1070.10	0.40
Std	8.16	14.68	3.38	19.78	1528.24	0.49
Min	18.00	0.00	1.00	0.00	0.00	0.00
25%	19.00	3.00	1.00	6.00	450.00	0.00
50%	21.50	8.00	1.00	13.00	450.00	0.00
75%	24.00	20.00	2.00	900.00	900.00	1.00
Max	64.00	80.00	54.00	163.00	24300.00	1.00

4.2.Feature Selection

Two methods were used for feature selection. The Light GBM Embedded feature selection method was used as well as correlation. The feature importance scores generated by Light GBM ranked months since last donation as the most important factor in predicting whether a donor will return to donate, it was followed by previous donations, age and months since first donation respectively. Blood group, gender, education and total volume donated had the least feature importance scores as shown in Figure 3 below.

Figure 3. Feature importance based on Light GBM



The feature importance based on correlation showed that months since last donation, Months since first donation, age, number of previous donations and total volume donated respectively as the most important features. It also showed that the number of previous donations and the total volume donated were highly correlated.

Table 2. Feature importance scores based on correlation

Feature	Correlation with 'Donated Blood in 2024'
Donated Blood in 2024	1.00000
Blood Group B+	0.02617
Blood Group O+	0.01563
Blood Group A+	0.01259
Blood Group A-	-0.001258

Gender	-0.02937
Blood Group O-	-0.04037
Blood Group B-	-0.06105
Education	-0.08243
Blood Group AB+	-0.12413
Previous Donations	0.13312
Total Volume Donated	-0.11313
Age	-0.15847
Months since First Donation	-0.55294
Months since Last Donation	-0.64021

4.3. Model Performance results

The model performed well on both training and testing data. Figure 4 below show the model confusion matrix.

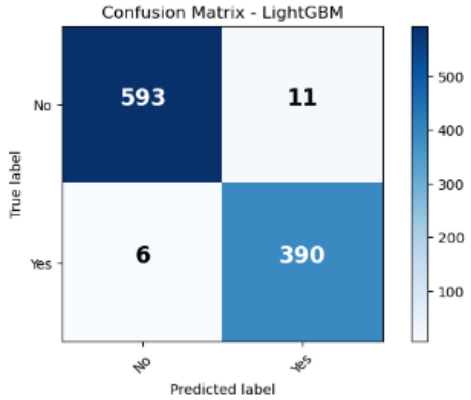


Figure 4. Light GBM confusion Matrix

The light GBM model achieved a performance accuracy of 0.9830, precision of 0.9726, a recall 0.9848 and F1 score of 0.9787 as shown in figure 5 below

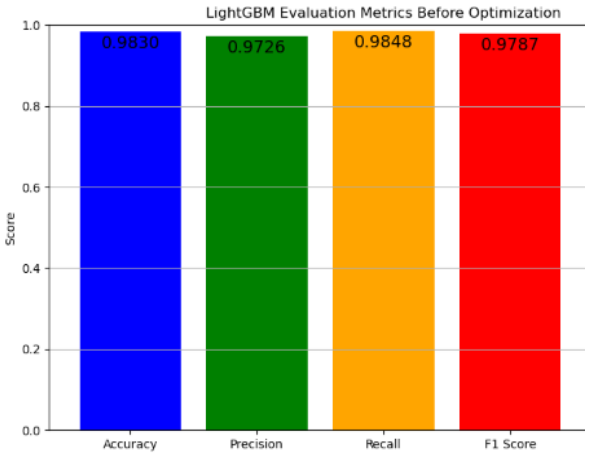


Figure 5. Light GBM Evaluation Metrics

Light GBM learning curve

The learning curve for the Light GBM models shows a considerable increase in accuracy on both training and cross validation sets as the number of iterations increase. This shows that the model is able to learn effectively as the amount of data increases. The model is able to learn the underlying patterns and generalize from the training data hence able to improve the performance. Figure 6. shows the light GBM accuracy learning curve.

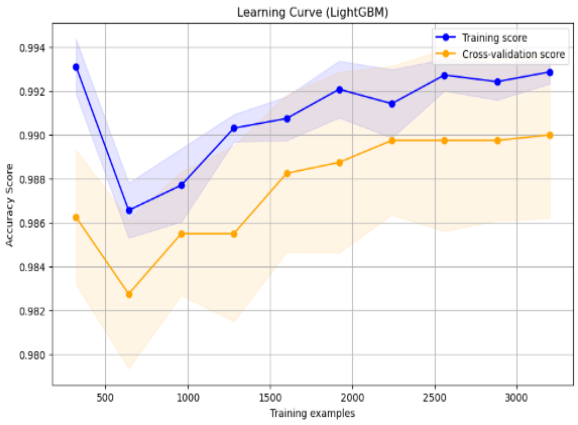


Figure 6. Light GBM learning curve

4.4.Comparative analysis of Light GBM with the existing models

Table 3. Comparative analysis of the Light GBM model with some of the existing models

Study	Algorithms	Dataset	Accuracy	F1 Score
[6]	Decision tree, and logistic regression.	748 donors with 5 variables	60	-
[7]	KNN	488 19 features	70.3	75.6
[5]	Random forest	81986 donors with 11 variables	91	-
[17]	Decision Treee C4.5	197 donors with Seven variables	84.17	-
[14]	SVM	748 donors	78.4	-
Light GBM model	**Light GBM**	**5000 9 features**	**98.3**	**97.8**

When compared to the existing blood donor retention models. The hybrid ensemble model achieved the highest accuracy of 98.3%. The model also achieved the best F1 score of 97.8% as compared to the existing studies. This highlights its superiority in identifying both the positive and the negative class.

It is important to note that most of the existing studies predominantly relied on accuracy as the main performance metric. However, accuracy can be misleading, particularly in imbalanced datasets where the majority class can dominate the

results in contrast, the F1 score provides a more balanced evaluation

5. CONCLUSION AND FUTURE WORK

This study developed and validated a gradient boosting model for blood donor retention based on the LightGBM. The results demonstrate that the Light GBM model achieved a considerably high performance across multiple evaluation metrics, including accuracy of 0.9830, precision of 0.9726, recall 0.9848 and F1 score 0.9787. This demonstrates its effectiveness and potential as a reliable tool for predicting whether a donor is likely to return to donate blood. This study contributes to the field of computer science by demonstrating how advanced machine algorithms can be employed to solve real world problems in healthcare.

Future studies could investigate the use of other machine learning algorithms, such as deep learning or reinforcement learning. Ensemble techniques can also be experimented and the performance be assessed to compare their effectiveness with the current Light GBM model for blood donor retention.

6. ACKNOWLEDGMENTS

We would like to express our heartfelt gratitude to all individuals and organizations who contributed to the successful completion of this study. Special thanks to the technical and administrative staff of Meru University of Science and Technology for their assistance and encouragement.

7. REFERENCES

[1] A. Van Dongen, “Easy come, easy go. Retention of blood donors,” Transfusion Medicine, vol. 25, no. 4, pp. 227–233, Aug. 2015, doi: 10.1111/tme.12249.

[2] J. Ou-Yang, C.-H. Bei, B. He, and X. Rong, “Factors influencing blood donation: a cross-sectional survey in Guangzhou, China,” Transfusion Medicine, vol. 27, no. 4, pp. 256–267, Aug. 2017, doi: 10.1111/tme.12410.

[3] World Bank, “Ensuring Access to Safe Blood in Kenya Amid COVID-19 Pandemic.” Accessed: Jul. 08, 2023. [Online]. Available: https://www.worldbank.org/en/news/feature/2022/05/06/ensuring-access-to-safe-blood-in-kenya-enhanced-amid-covid-19-pandemic

[4] A. Panesar, Machine Learning and AI for Healthcare. Berkeley, CA: Apress, 2019. doi: 10.1007/978-1-4842-3799-1.

[5] M. Cloutier, Y. Grégoire, K. Choucha, A. Amja, and A. Lewin, “Prediction of donation return rate in young donors using machine-learning models,” ISBT Sci Ser, vol. 16, no. 1, pp. 119–126, Feb. 2021, doi: 10.1111/voxs.12618.

[6] C. Marade, A. Pradeep, D. Mohanty, and C. Patil, “Forecasting Blood Donor Response Using Predictive Modelling Approach,” International Journal of Computer Science and Mobile Computing, vol. 8, no. 4, pp. 73–77, 2019.

[7] K. Pabreja and A. Bhasin, “A Predictive Analytics Framework for Blood Donor Classification,” International Journal of Big Data and Analytics in Healthcare, vol. 6, no. 2, pp. 1–14, Jul. 2021, doi: 10.4018/IJBDAH.20210701.oa1.

[8] N. H. A. Malek, W. F. W. Yaacob, Y. B. Wah, S. A. Md Nasir, N. Shaadan, and S. W. Indratno, “Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data,” Indonesian Journal of Electrical Engineering and Computer Science, vol. 29, no. 1, p. 598, Jan. 2022, doi: 10.11591/ijeecs.v29.i1.pp598-608.

[9] Y. Dei-Adomakoh, L. Asamoah-Akuoko, B. Appiah, A. Yawson, and E. Olayemi, “Safe blood supply in sub-Saharan Africa: challenges and opportunities,” Lancet Haematol, vol. 8, no. 10, pp. e770–e776, Oct. 2021, doi: 10.1016/S2352-3026(21)00209-X.

[10] S. Liu et al., “Machine learning models to predict red blood cell transfusion in patients undergoing mitral valve surgery,” Ann Transl Med, vol. 9, no. 7, pp. 530–530, Apr. 2021, doi: 10.21037/atm-20-7375.

[11] World Health Organization, “Global status report on blood safety and availability,” 2017.

[12] S. Campagnini, C. Arienti, M. Patrini, P. Liuzzi, A. Mannini, and M. C. Carrozza, “Machine learning methods for functional recovery prediction and prognosis in post-stroke rehabilitation: a systematic review,” J Neuroeng Rehabil, vol. 19, no. 1, p. 54, Dec. 2022, doi: 10.1186/s12984-022-01032-4.

[13] W. B. Zulfikar, Y. A. Gerhana, and A. F. Rahmania, “An Approach to Classify Eligibility Blood Donors Using Decision Tree and Naive Bayes Classifier,” in 2018 6th International Conference on Cyber and IT Service Management (CITSM), IEEE, Aug. 2018, pp. 1–5. doi: 10.1109/CITSM.2018.8674353.

[14] P. Selvaraj, A. Sarin, and B. I. Seraphim, “Forecasting System for Donation of Blood Using SVM Model,” Int J Res Appl Sci Eng Technol, vol. 10, no. 5, pp. 136–140, May 2022, doi: 10.22214/ijraset.2022.41940.

[15] T. R. Mahesh et al., “AdaBoost Ensemble Methods Using K-Fold Cross Validation for Survivability with the Early Detection of Heart Disease,” Comput Intell Neurosci, vol. 2022, pp. 1–11, Apr. 2022, doi: 10.1155/2022/9005278.

[16] S. Bates, T. Hastie, and R. Tibshirani, “Cross-Validation: What Does It Estimate and How Well Does It Do It?,” J Am Stat Assoc, pp. 1–12, May 2023, doi: 10.1080/01621459.2023.2197686.

[17] C. Salazar-Concha and P. Ramírez-Correa, “Predicting the Intention to Donate Blood among Blood Donors Using a Decision Tree Algorithm,” Symmetry (Basel), vol. 13, no. 8, p. 1460, Aug. 2021, doi: 10.3390/sym13081460.