Will They Donate?

By Anusha Kankipati and Anagha Nair

Abstract

This project aims to predict whether a blood donor will make a donation during a specific period, focusing on March 2007. Using data from the Blood Transfusion Service Center in Hsin-Chu City, Taiwan, the dataset includes four primary features. To enhance predictive performance, we augmented the dataset by engineering additional features derived from the original data, offering deeper insights into donor behavior. The target variable is a binary outcome indicating whether a donation occurred in the specified period.

We explore three classification methods: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Logistic Regression. These methods were implemented manually in Python, enabling a comprehensive understanding of their theoretical foundations and practical applications. The dataset was split into training and testing subsets to evaluate model performance. In addition to measuring accuracy, we constructed a confusion matrix to analyze true positives, true negatives, false positives, and false negatives, providing a detailed assessment of each model's effectiveness.

LDA and QDA were used to evaluate class separability, while Logistic Regression served as a robust baseline for comparison. Feature augmentation improved the models' ability to capture complex patterns in the data, enhancing their predictive capabilities. Due to the imbalanced dataset, attempts were made to avoid underpredicting of the majority category.

1 Introduction

This year, the American Red Cross declared a national blood shortage. Blood transfusions are life-saving procedures, with each donation potentially saving up to three lives. Despite someone in the United States needing blood every two seconds, only 3% of eligible individuals donate. This study aims to identify factors that correlate with blood donors, particularly focusing on what motivates individuals to donate repeatedly. [1] Understanding these factors is critical in addressing the shortage and could help save countless lives.

In this paper, the authors analyze data from the Blood Transfusion Service Center in Hsin-004 Chu City, Taiwan to see whether the features given (Months since Last Donation, Number of Donations, Total Volume Donated in c.c., and Months since First Donation) could be used to predict whether somebody would donate again during a certain time period, specifically March 2007. Since this is a binary classification problem (donation or no donation), the authors chose three methods: Logistic Regression, Linear Discriminant Analysis, and Quadratic Discriminant Analysis. An accuracy score and confusion matrix will be computed for each method in order to determine which is the most effective in this task.

2 Background/Related Work

Predicting blood donor behavior is a critical challenge for blood centers striving to maintain a stable supply chain while optimizing outreach efforts. A study by Kauten et al. (2021) from Auburn University [2] investigates this problem using operational data from a large U.S. blood center. The research explores the application of machine learning models to predict donor retention and inform cost-effective outreach strategies.

The authors developed a comprehensive pipeline for processing relational donor data into tabular datasets suitable for machine learning. The study utilized advanced algorithms, including Random Forest (RF) and Gradient Boosting (GB), achieving impressive results with an 80% sensitivity and 99% specificity. These models were tuned using strati-

[1] https://www.hcahoustonhealthcare.com/healthy-living/blog/donating-blood-benefits-both-recipients-and-donors

[2] https://www.eng.auburn.edu/~xqin/pubs/predicting_blood_kauten_isf21.pdf

fied cross-validation and evolutionary algorithms to optimize hyperparameters. The evaluation employed metrics like accuracy, sensitivity, specificity, and Matthews Correlation Coefficient (MCC), highlighting RF as the most effective model.

Kauten et al.'s research highlights the potential of data-driven approaches to enhance blood center operations by targeting likely donors and reducing unnecessary outreach. Their work complements prior studies, including applications of neural networks, support vector machines, and decision trees, providing a robust foundation for exploring predictive analytics in blood donation systems. This study underscores the importance of advanced machine learning techniques in addressing public health challenges related to blood supply management.

3 Approach

For this project, we decided to use various classification models, such as LDA, QDA, and Logistic Regression to determine whether or not someone donated blood in March 2007. We wanted compare the training and test accuracy of all 3 methods, as well as see which produced the most ideal ROC curve and confusion matrix. We also noticed collinearity in the data, which was affecting the performance of LDA and QDA because there was repeated data. Therefore we reduced the set to 2 dimensions to produce better results.

3.1 Principal Component Analysis (PCA)

PCA identifies the directions (principal components) that maximize variance in the data. This is achieved by solving the eigenvalue problem:

$$\Sigma v = \lambda v$$

where:

- Σ is the covariance matrix of the data,
- v is the eigenvector (principal component),
- λ is the eigenvalue corresponding to v.

The steps to compute PCA are:

1. Compute the covariance matrix Σ from the dataset X.
2. Solve the eigenvalue equation $\Sigma v = \lambda v$.
3. Select the eigenvectors $v_1, v_2, \ldots, v_k$ corresponding to the k largest eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$.
4. Project the data X onto these eigenvectors to obtain the principal components:

$$Z = XV_k$$

where V_k is the matrix of the top k eigenvectors, and Z is the transformed dataset in the new principal component space.

3.2 Logistic Regression

Logistic regression predicts the probability $P(y = 1 \mid x)$ of a binary outcome given the feature vector x. The probability is modeled using the sigmoid function:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-z}}, \quad z = w^\top x + b$$

where:

- z is the linear combination of input features x,
- $w \in R^d$ is the weight vector,
- $b \in R$ is the intercept term.

3.3 Linear Discriminant Analysis (LDA)

LDA assumes that the features for each class $y \in \{0, 1\}$ follow a Gaussian distribution with a shared covariance matrix Σ. The class-conditional densities are:

$$P(x \mid y = 0) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_0)^\top \Sigma^{-1}(x-\mu_0)}$$

$$P(x \mid y = 1) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_1)^\top \Sigma^{-1}(x-\mu_1)}$$

In the decision boundary quadratic equation, here is no quadratic term and only a linear term, as each class shared the same covariance matrix.

3.4 Quadratic Discriminant Analysis (QDA)

QDA assumes that features for each class $y \in \{0, 1\}$ follow a Gaussian distribution with class-specific covariance matrices Σ_0 and Σ_1. The class-conditional densities are:

$$P(x \mid y = 0) = \frac{1}{(2\pi)^{d/2}|\Sigma_0|^{1/2}} e^{-\frac{1}{2}(x-\mu_0)^\top \Sigma_0^{-1}(x-\mu_0)}$$

$$P(x \mid y = 1) = \frac{1}{(2\pi)^{d/2}|\Sigma_1|^{1/2}} e^{-\frac{1}{2}(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1)}$$

In the decision boundary quadratic equation, the quadratic terms in arise due to the class-specific covariance matrices.

4 Experiments

Below are outlined the dataset, experiment, evaluation, and result specs of our project.

4.1 Dataset

We used the **Blood Transfusion Service Center Dataset** [3] from the UCI Machine Learning Repository. It contains 4 features:

- Months since Last Donation: this is the number of months since this donor's most recent donation.
- Number of Donations: this is the total number of donations that the donor has made.
- Total Volume Donated: this is the total amount of blood that the donor has donated in cubic centimeters.
- Months since First Donation: this is the number of months since the donor's first donation.

However, this was a relatively sparse dataset to work with, so we decided to add some other features. We added 4 new features:

- Feature 1 = 0.5×Recency (months) + 0.3×Frequency (times) - 0.2×Time (months)
- Feature 2 = 0.7×Frequency (times) - 0.4×Monetary (c.c. blood) + 0.2× Time (months)
- Feature 3 = 0.6×Recency (months) + 0.5×Monetary (c.c. blood) - 0.3×Frequency (times)
- Feature 4 = -0.5×Recency (months) + 0.4×Time (months) + 0.3×Frequency (times)

Notably, the dataset is imbalanced, with about 90% of data instances being in the negative category. The minority category was about 10%, which was the proportion of donors who donated again.

4.2 How we ran our experiments

We first loaded the data, and split the dataset into 80% training and 20% testing. We then used Logistic regression to test the accuracy of the model on out train and test datasets. For LDA and QDA, we had to use PCA to remove collinearity. We then calculated the accuracy of both models on the test and train datasets. All the models took took almost an instantanous time to run, as the dataset is not relatively that large.

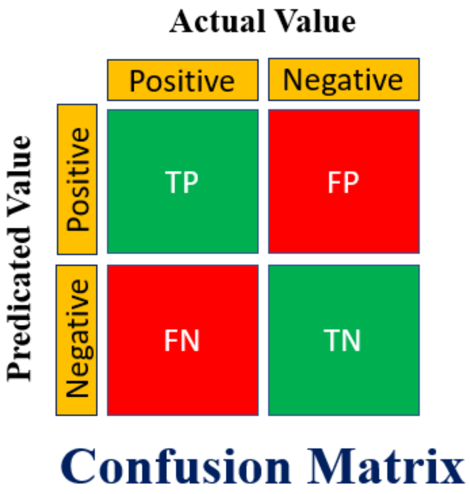


Figure 1: TechPlayon Example [4]

4.3 Evaluation Metrics

We used two main evaluation metrics to measure the efficacy of our methods: accuracy scores and confusion matrices. An accuracy score will measure the percent of correct predictions versus the number of total samples. The score will be between 0 and 1. A score of 1 is perfect score, indicating that the model is completely accurate for the provided data. Accuracy score can be calculated using this equation:

$$Accuracy = \frac{\#ofcorrectpredictions}{total\#ofpredictions}$$

A confusion matrix will visually demonstrate the accuracy of the model by showing how many samples fit into 4 categories: false positive, false negative, true positive, and true negative. As illustrated in Figure 1, each quandrant represents a different one of these categories. False negatives and false positives both reduce accuracy, while true positives and negatives increase the accuracy score.

Using the confusion matrix, we were also able to consider other metrics, such as precision and recall which are also good indicators of the performance of the model, especially in the context of imbalanced datasets. These metrics consider the ratio of true positives to false positives and false negatives respectively. These two metrics can be combined into an F score, as shown below.

$$Precision = \frac{\#TP}{\#TP + \#FP}$$

[3] https://archive.ics.uci.edu/dataset/176/blood+transfusion+service+center

$$Recall = \frac{\#TP}{\#TP + \#FN}$$

$$F = \frac{\#2PR}{P + R}$$

4.4 Results

Here are our results for each model:

Before Augmenting the Dataset

- **Logistic Regression:** Test Accuracy = 0.9067
- **LDA:**
  - Training Accuracy = 0.7241
  - Test Accuracy = 0.9000
- **QDA:**
  - Training Accuracy = 0.7258
  - Test Accuracy = 0.9067

After Augmenting the Dataset

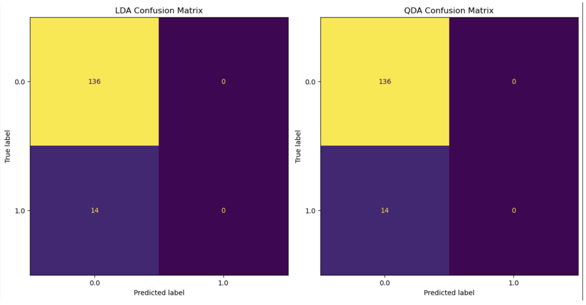
- **LDA:**
  - Training Accuracy = 0.7258
  - Test Accuracy = 0.9067
- **QDA:**
  - Training Accuracy = 0.7291
  - Test Accuracy = 0.8933

If the dataset is not normalized, the QDA test accuracy decreases to 0.886.

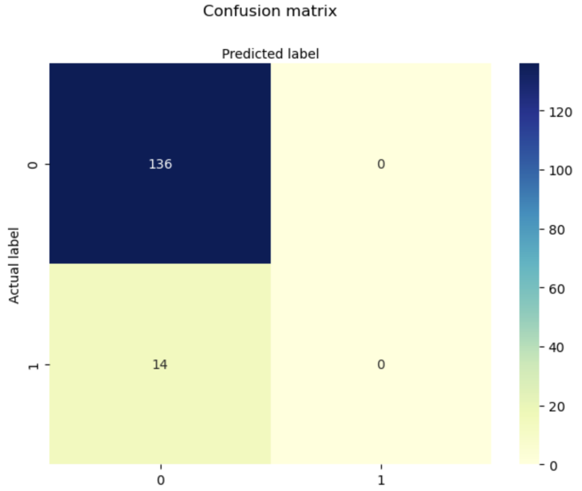
What is interesting is that the confusion matrix also stayed the same for each implementation, as shown below:

Confusion Matrices

LDA and QDA Confusion Matrices:



Logistic Regression Confusion Matrix:



We noticed is an issue where we are only getting true negatives and false negatives, as we can see in the confusion matrix. The issue is that the models are predicting all instances as "false" meaning that it is the majority case. Therefore, the fact that our accuracy is relatively high is misleading. Since the majority category is the negative category, we decided to flip the positive and negative when considering precision and recall. We found that our results gave

$$R = \frac{136}{136 + 0} = 1$$

$$R = \frac{136}{136 + 14} = 0.91$$

$$F = \frac{2(1)(.91}{1 + .91} = 0.95$$

So mitigate this, we tried implementing focal loss, as well as changing the weights for gradient descent for each class in logistic regression. However, this did not yield good results or improve accuracy. We did note that the optimal weight ratio between the majority and minority class was 1 to 1.4. We conclude that this could potentially be due to not enough training data to train the model to reliably make predictions.

5 Conclusion

We learned that Logistic Regression, LDA and QDA all perform relatively the same on this dataset. PCA is necessary to remove collinearity, and normalizing the dataset help improve accuracy in some cases. We noticed that the model's main issue is overpredicting the majority category and underpredicting the minority category. While the accuracy remained relatively high, the precision and recall values and therefore low. The few methods we attempted to mitigate this did not yield good results, suggesting that more advanced strategies may need to be imployed. In the future, we could use more types of models to see if that affects accuracy, such as cross-validation or adding gradient descent to the logistic regression model.

References

[1] HCA Houston Healthcare. (n.d.). *Donating blood benefits both recipients and donors*. https://www.hcahoustonhealthcare.com/healthy-living/blog/donating-blood-benefits-both-recipients-and-donors

[2] Kauten, C., Gupta, A., Qin, X., and Richey, G. (2021). *Predicting blood donors using machine learning techniques*. https://www.eng.auburn.edu/~xqin/pubs/predicting_blood_kauten_isf21.pdf

[3] Yeh, I. C., Yang, K. J., and Ting, T. M. (2009). *Blood transfusion service center dataset*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/176/blood+transfusion+service+center

[4] DrivenData. (n.d.). *Warm-up: Predict blood donations*. https://www.drivendata.org/competitions/2/warm-up-predict-blood-donations/page/7/

[5] TechPlayon. (n.d.). *Understanding confusion matrix in machine learning*. https://www.techplayon.com/understanding-confusion-matrix-in-machine-learning/