**1. Introduction**

The main aim of this project is to predict the donor will donate the blood in March 2007. Blood transfusion is a medical procedure that saves millions of lives every year. Maintaining an adequate blood supply is challenging, and predicting donor retention can significantly enhance the efficiency of blood donation drives. This report presents an in-depth analysis of blood donation data, compares multiple predictive models, and discusses the challenges encountered during the analysis.

**2. Data Analysis**

**2.1 Dataset Overview**

The dataset consists of blood donation records from a mobile blood donation vehicle in Taiwan. Below is the dataset to predict the donor will donate the blood in March 2007.

* **Months since Last Donation**: Time since the donor's most recent donation.
* **Number of Donations**: Total count of donations made by the donor.
* **Total Volume Donated**: Total amount of blood donated in cubic centimetres.
* **Months since First Donation**: Time since the donor's first-ever donation.
* **Made Donation in March 2007**: Binary target variable indicating whether the donor contributed in March 2007.

**2.2 Exploratory Data Analysis (EDA)**

* **Distribution Analysis**: Histograms was used to visualize the spread of each numerical feature.
* **Correlation Analysis**: The Pearson correlation coefficient was computed to assess relationships between features.
* **Missing Values**: No missing values and outliers were detected in the dataset.
* **Feature Importance**: The "Volume of donated blood in c.c" and "Months since Last Donation" appeared to be the most influential predictors.

**3. Predictive Model Comparison**

**3.1 Model Selection**

Implemented various models to determine the donor retention:

* **Logistic Regression**
* **Random Forest Classifier**
* **Gradient Boosting Classifier**

**3.2 Model Evaluation**

* Performance measured using:
  + Root Mean Squared Error (RMSE)
  + Mean Squared Error (MSE)
  + R-squared Score
* Best performing model: XG Boosting model shows the highest accuracy point.

**3.3 Performance Metrics**

Models were assessed using:

* **Accuracy**
* **Precision, Recall, F1-score**, **and Support**

**3.4 Results & Best Model Selection**

Below table predictions values are shown by applying log normalization method.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **Support** |
| Logistic Regression | 0.767 | 0-0.79 1-0.56 | 0-0.95 1-0.18 | 0-0.86 1-0.27 | 0-88 1-28 |
| Random Forest | 0.707 | 0-0.78 1-0.33 | 0-0.86 1-0.21 | 0-0.82 1-0.26 | 0-88 1-28 |
| Gradient Boosting | 0.690 | 0-0.77 1-0.30 | 0-0.84 1-0.21 | 0-0.80 1-0.25 | 0-88 1-28 |

Below table prediction values are shown by applying SMOTE method.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** | **Support** |
| Logistic Regression | 0.699 | 0-0.74 | 0-0.63 | 0-0.68 | 0-89 |
| 1-0.67 | 1-0.77 | 1-0.72 | 1-87 |
| Random Forest | 0.761 | 0-0.75 | 0-0.79 | 0-0.77 | 0-89 |
| 1-0.77 | 1-0.74 | 1-0.75 | 1-87 |
| Gradient Boosting | 0.79 | 0-0.77 | 0-0.83 | 0-0.80 | 0-89 |
| 1-0.81 | 1-0.75 | 1-0.78 | 1-87 |

From the comparison, the **Gradient Boosting model** was selected based on the highest Accuracy score and F1-score by using smote technique.

**4. Challenges and Solutions**

**4.1 Challenges Encountered**

* **Imbalanced Dataset**: The dataset had an unequal distribution of positive and negative samples.
* **Feature Scaling**: Some models required normalization for optimal performance.
* **Overfitting**: Certain complex models showed high accuracy on training data but underperformed on test data.

**4.2 Techniques Used to Overcome Challenges**

* **Resampling Techniques**: Used SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset.
* **Cross-Validation**: Implemented k-fold cross-validation to improve generalizability.
* **Hyperparameter Tuning**: Used GridSearchCV for optimizing model parameters.

**5. Conclusion & Recommendations**

Based on our analysis, the most effective model for predicting blood donor retention was **Gradient Boosting**, achieving the highest predictive accuracy and reliability. Now we can target the people who are interested in donating blood, and which will result in getting more volunteers and we can save more people.