Case Study

Using Machine Learning for Analysis of Blood Donation:
Which factors impact high donation rates?

머신러닝을 활용한 헌혈 참여 변수 분석

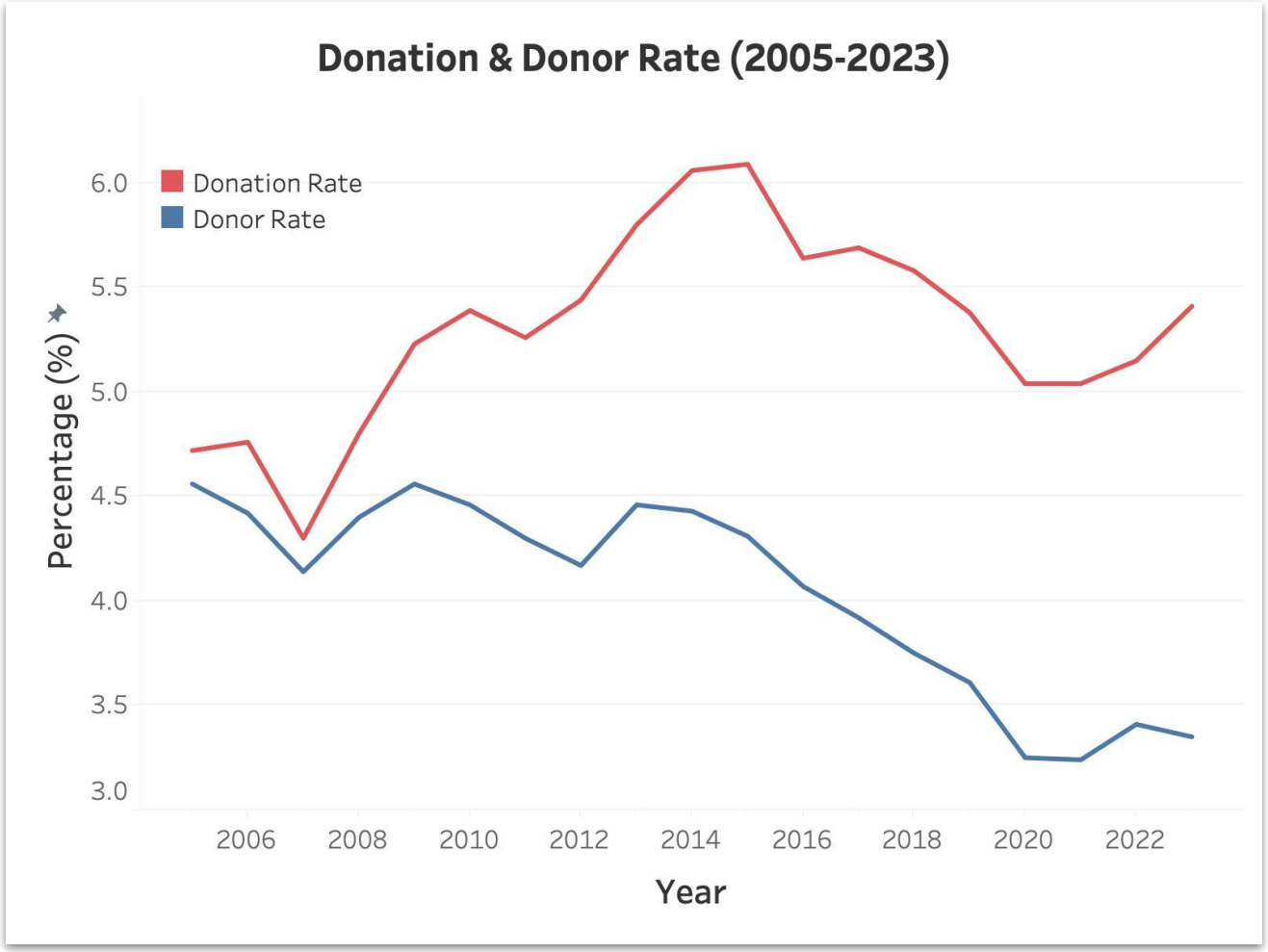
Hyun Jin (Austin) Kang, Yoojin (Audrey) Jung (Data Analyst)

Table of Contents

Background

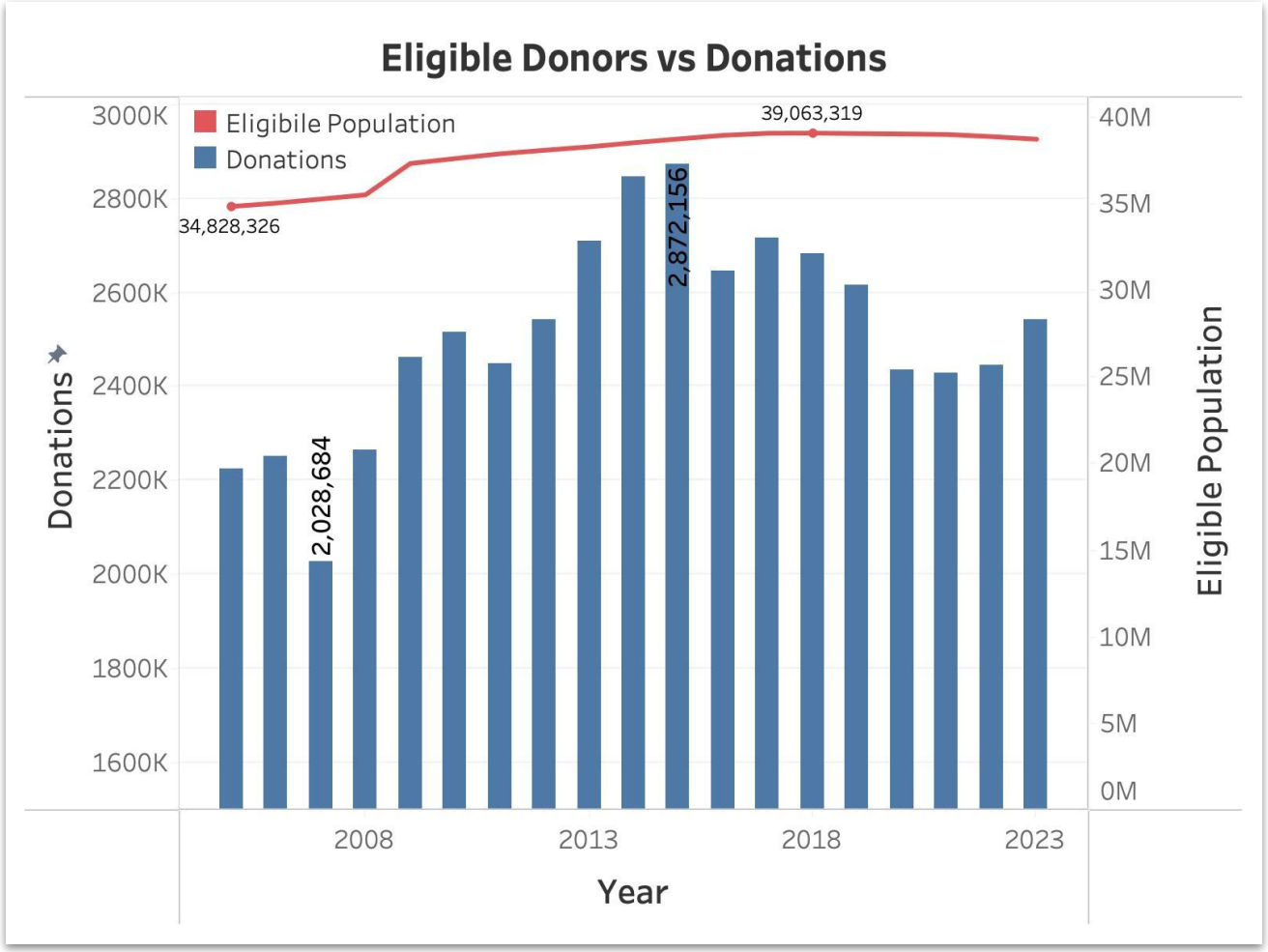
- Blood donation is the only method to securing accessibility of blood products for patients in need of transfusion
- Accessibility is solely dependent on voluntary blood donation
- Donation rate in South Korea maintains 5% after reaching 6% in 2015
- Donor rate* gradually decreases since it is first recorded in the Korean Statistical Information Service in 2005

** Donor rate refers to count of actual donors against the eligible population.*



Background

- Eligible population is steady after peaking in 2018 with approximately 39 million donors
- Donation count varies yearly and shows decreasing trend since peaking in 2015
- Total population in South Korea is currently declining adding challenges to maintaining the blood donation rate



Objective

- Analyze the trend of blood donations since 2005
- Identify which variables have an impact on high blood donation
- Recommendations for blood centers to increase blood donation

Pace : Plan Stage

Plan (Data Collection)

- Korean Statistical Information Service (KOSIS)
  - Blood information statistics
    - Donation by age group
    - Types of donation
    - Location of donation
  - Registered population statistics
    - Population by region

Plan (Data Dictionary)

Term	Definition
Blood Center	Korean Red Cross blood center (region of blood center)
Eligible population	Population between ages 16 and 69 (ages 16 to 65 prior to 2009)
Total donations	Blood donation count at Korean Red Cross blood centers
Whole blood donation	Type of donation that collects all components including red blood cell, platelet, plasma, and white cells
Apheresis donation	Type of donation that only collects specific components (ex. platelet, plasma, platelet-plasma)
Mobile donation	Donations that take place at mobile blood donation vehicles or separate prepared location
Fixed-site donation	Donations that take place at fixed blood donation centers
Age groups	16-19, 20-29, 30-39, 40-49, 50-59, 60-69

** This case study is based on blood donation from the Korean Red Cross*

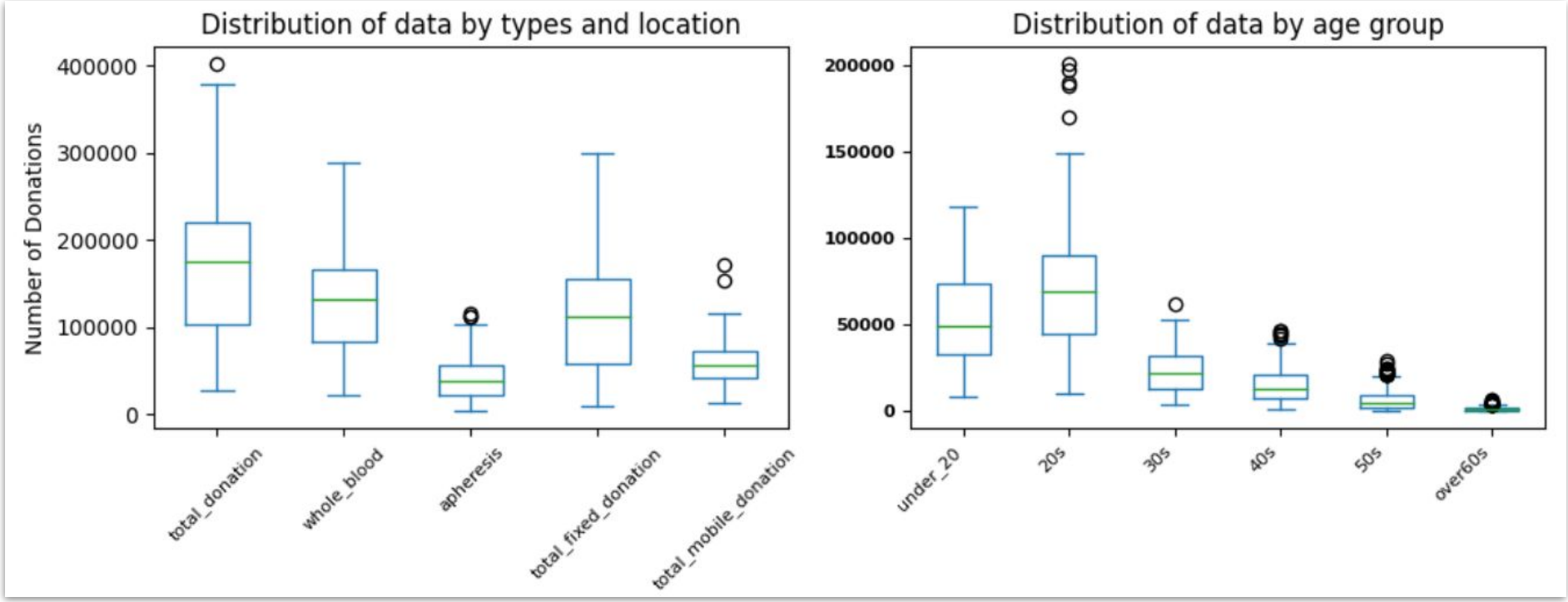
Plan - Preprocessing and feature engineering

- Evaluate data types are set appropriately with no duplicated or null data
- Create 'total donation' column by adding 'fixed-site donation' and 'total mobile donation' columns
- Merged columns that have low proportions
    - '50s' + 'over_60s' → 'over_50s'
    - 'mobile_religious_org' + 'mobile_other' → 'mobile_others'
- Set top 25% of total donation performance for each center as 'high_donation' (1 or 0)
- To get donation rate by year, merged 'eligible population' data with 'df_by_year'
  → Donation Rate = Total donations / Total eligible population
- To get contributions of each center for each year, made a column 'total_donation_prop'
  → Donation proportion = Total donations / National total donation
- To get donation rate by region, merged three seoul blood centers (seoul_central, seoul_nambu, seoul_dongbu)

pAce: Analyze Stage

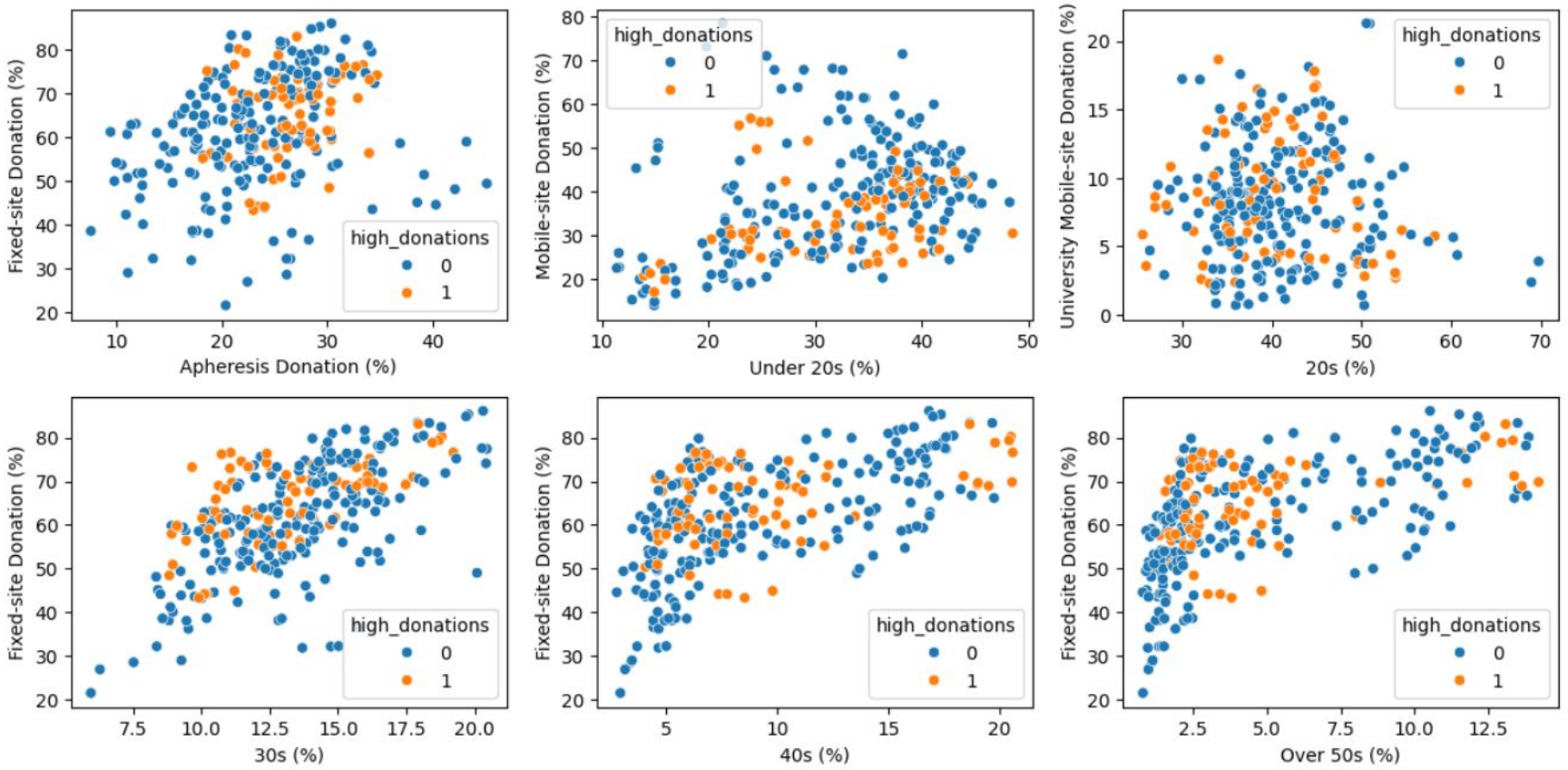
Data Distribution

- Identified few outliers in target variable (total donations)
  → Kept these outliers considering the characteristic of tree-based models
  → Scaled data before training the model to revise the outlier

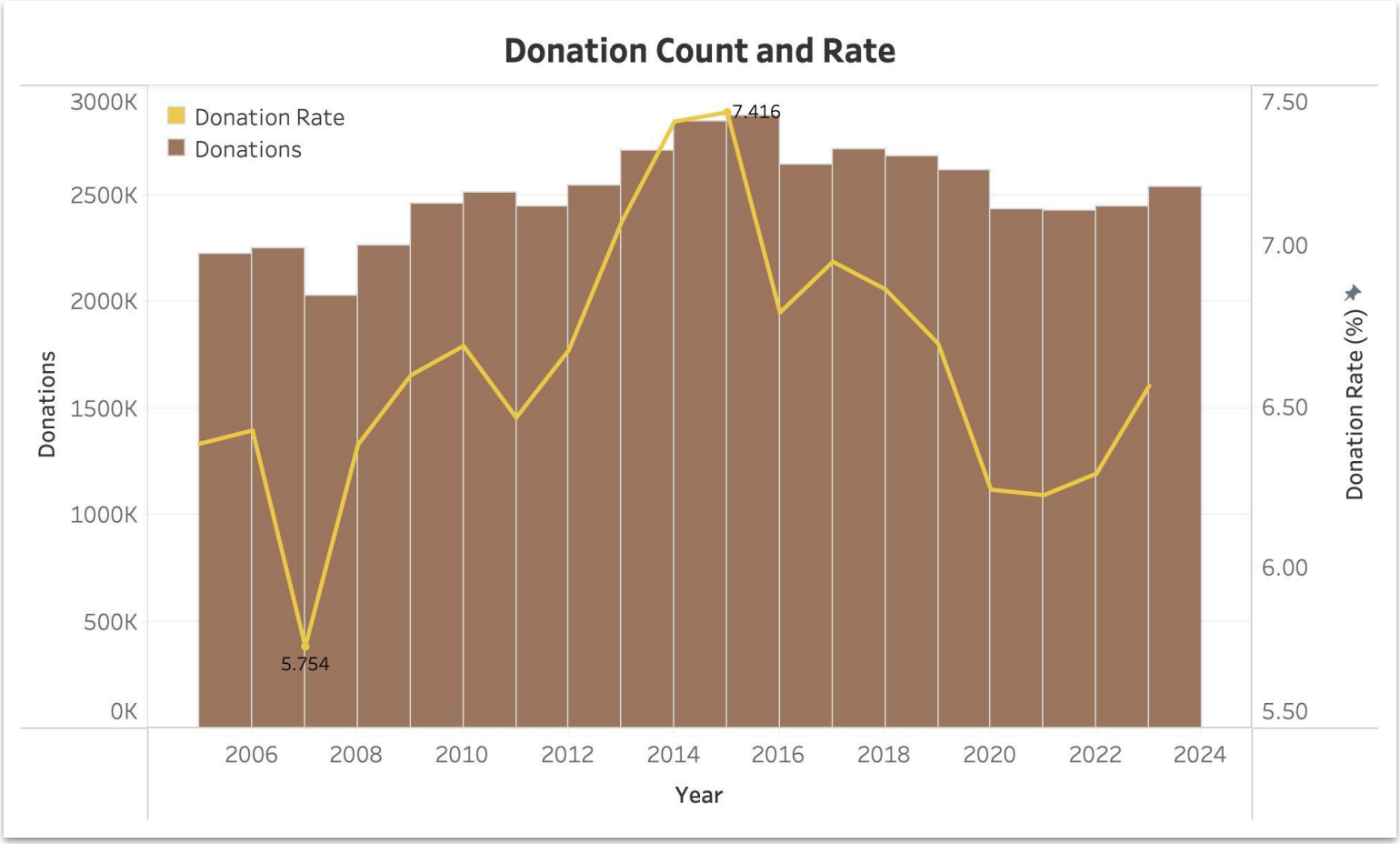


Relationship between features

- Identified some positive relationships between features.
- However, unable to identify which features impact high donations through plots.
  - This indicates that the high donations are unable to be explained properly through simple regression.



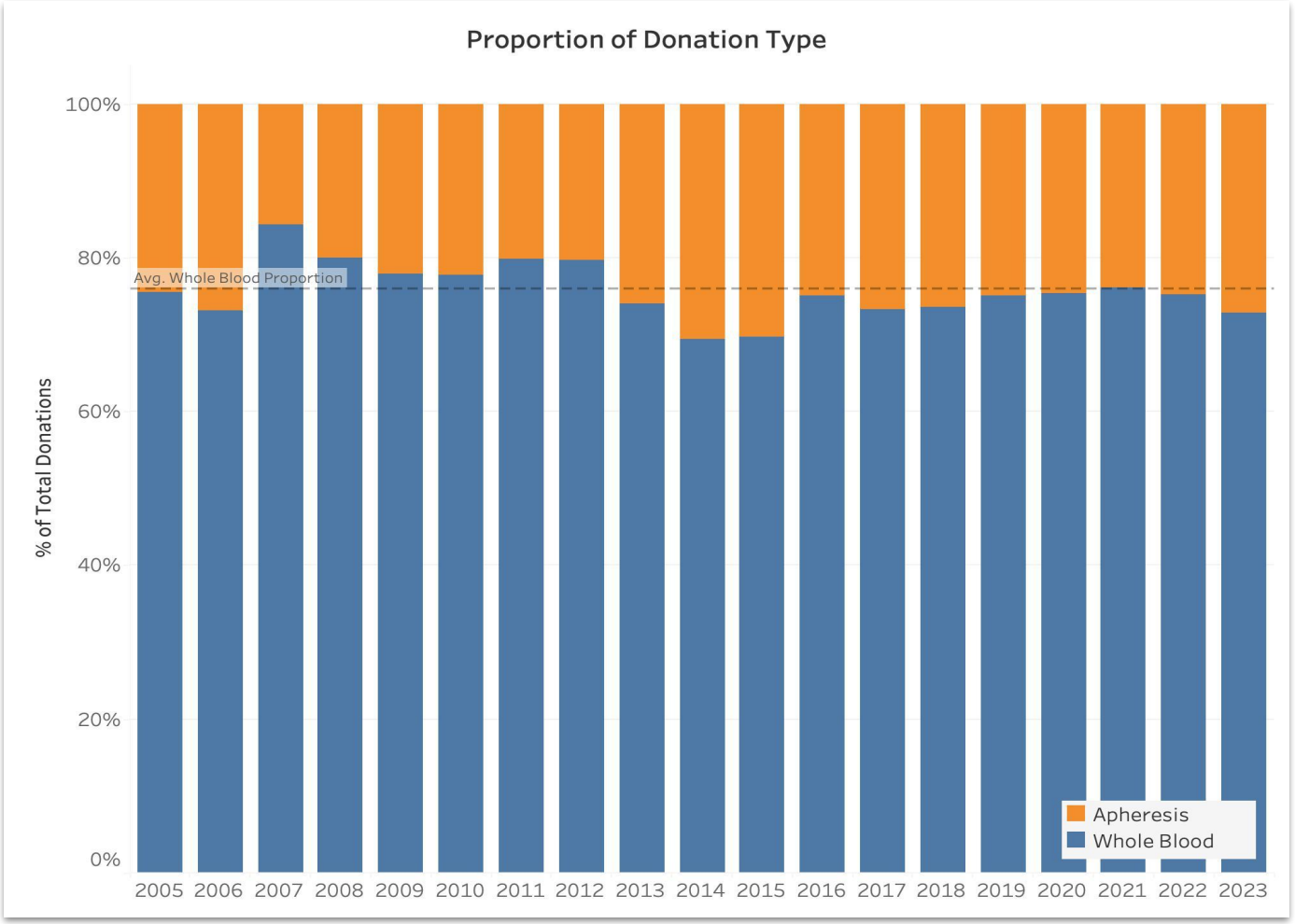
Exploratory Data Analysis (EDA) - by year



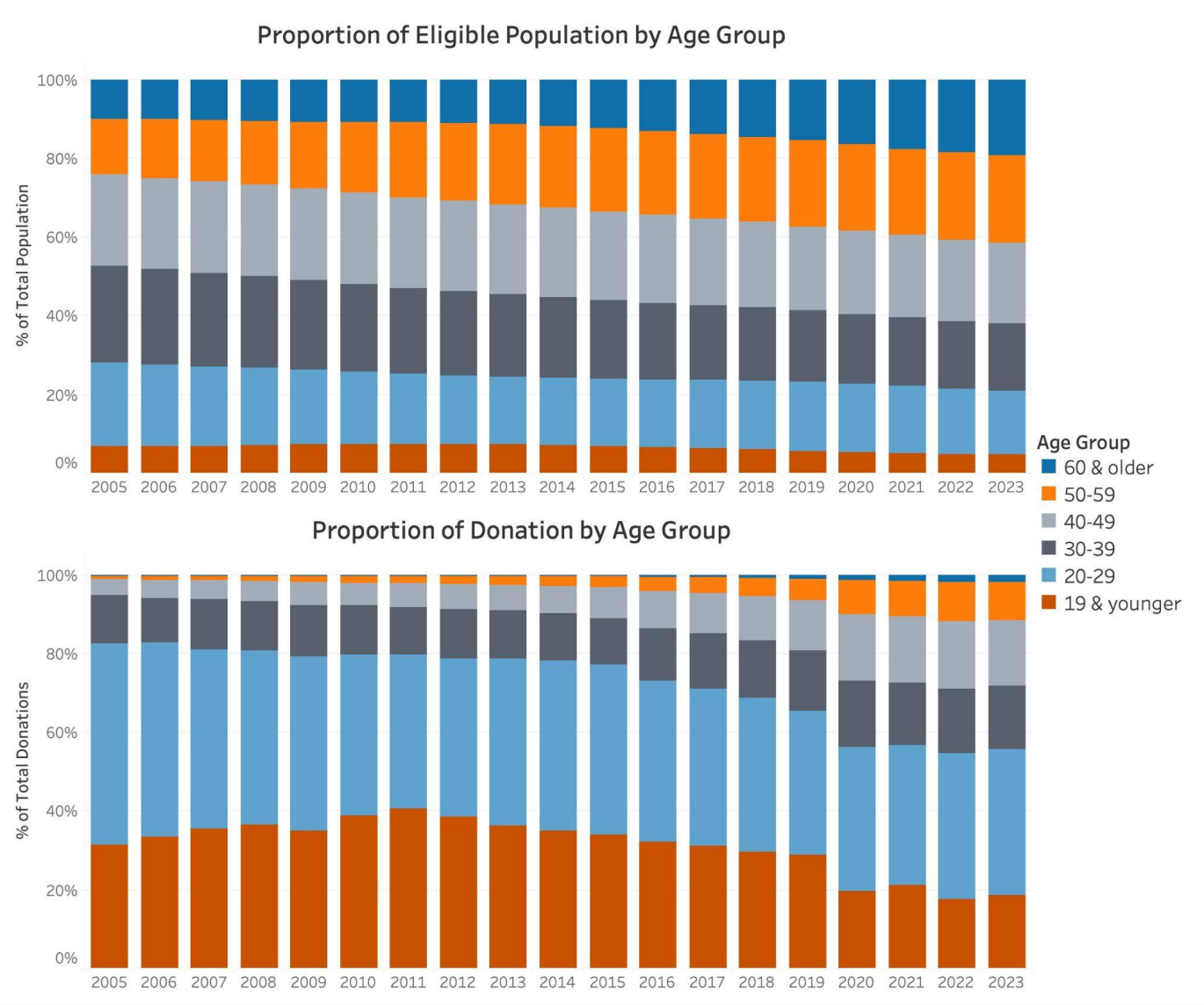
- Lowest donation count was in 2007 with approximately 2 million donations at a rate of 5.7% of the total eligible population
  - **Lower than during the COVID-19 pandemic at 6.2%**
- Highest donation count was in 2015 with approximately 2.8 million donations at a rate of 7.4% of the total eligible population
  - **Since then, the donation rate has shown a downward trend**
- Donation count is in the recovery stage following the COVID-19 pandemic

Exploratory Data Analysis (EDA) - by year

- No significant change has been identified in the proportion of donation type between whole blood or apheresis
- The average whole blood proportion is approximately 76%

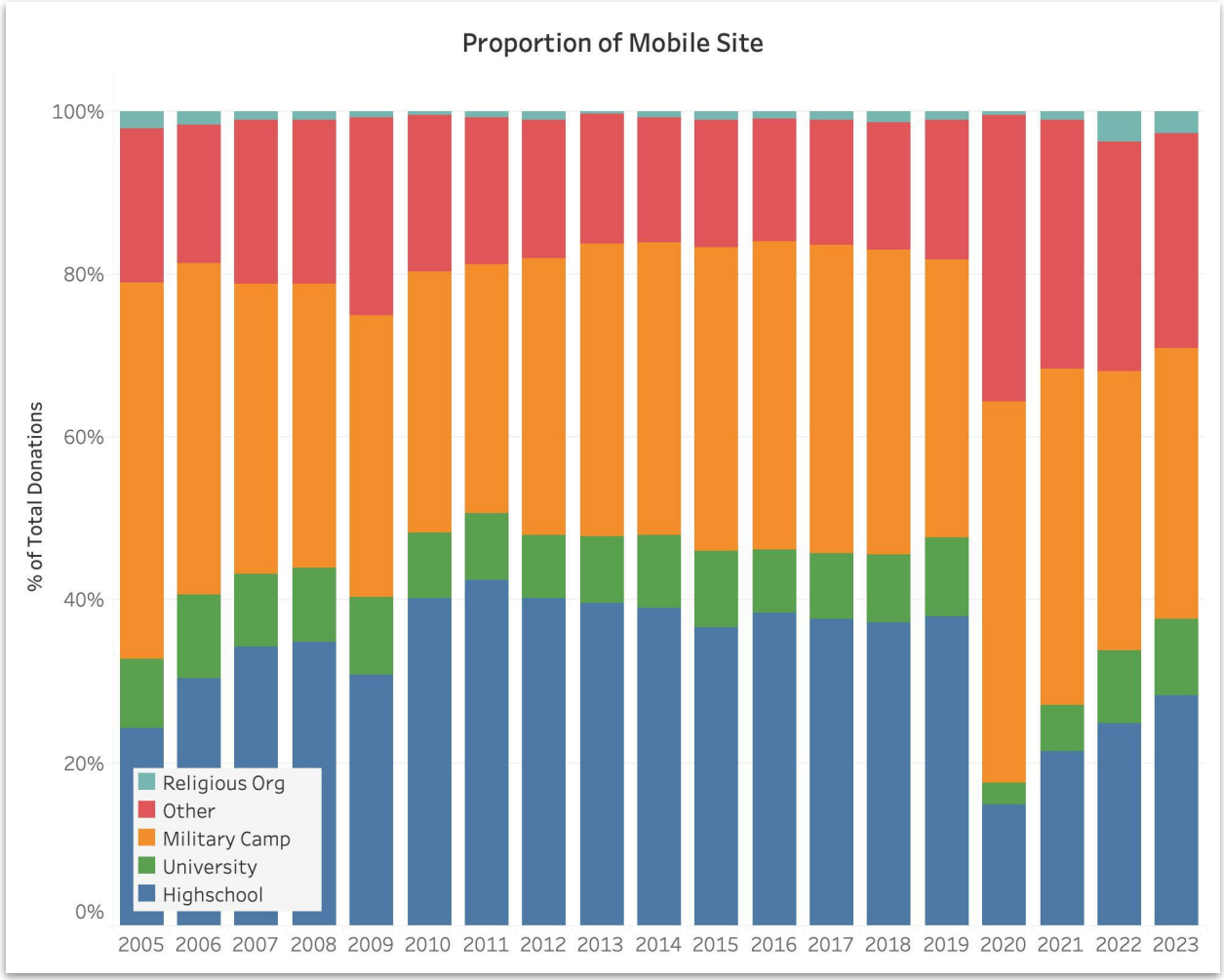


Exploratory Data Analysis (EDA) - by year/age



- There is an imbalanced proportion between the **eligible** population and **donor** population
- Youngest two age groups (16-19, 20-29) account for 25% of the eligible population but **contributes up to 70% of total donations**
- Oldest three age groups (40's, 50's, 60's) account for over 50% of the eligible population but **only contributes 14% of total donations**
- Note the eligible population of ages 50 and older is increasing every year due to the continued low birth rate and aging society of South Korea
- Also note there is a gradual decrease in donations from the youngest two age groups (16-19, 20-29)

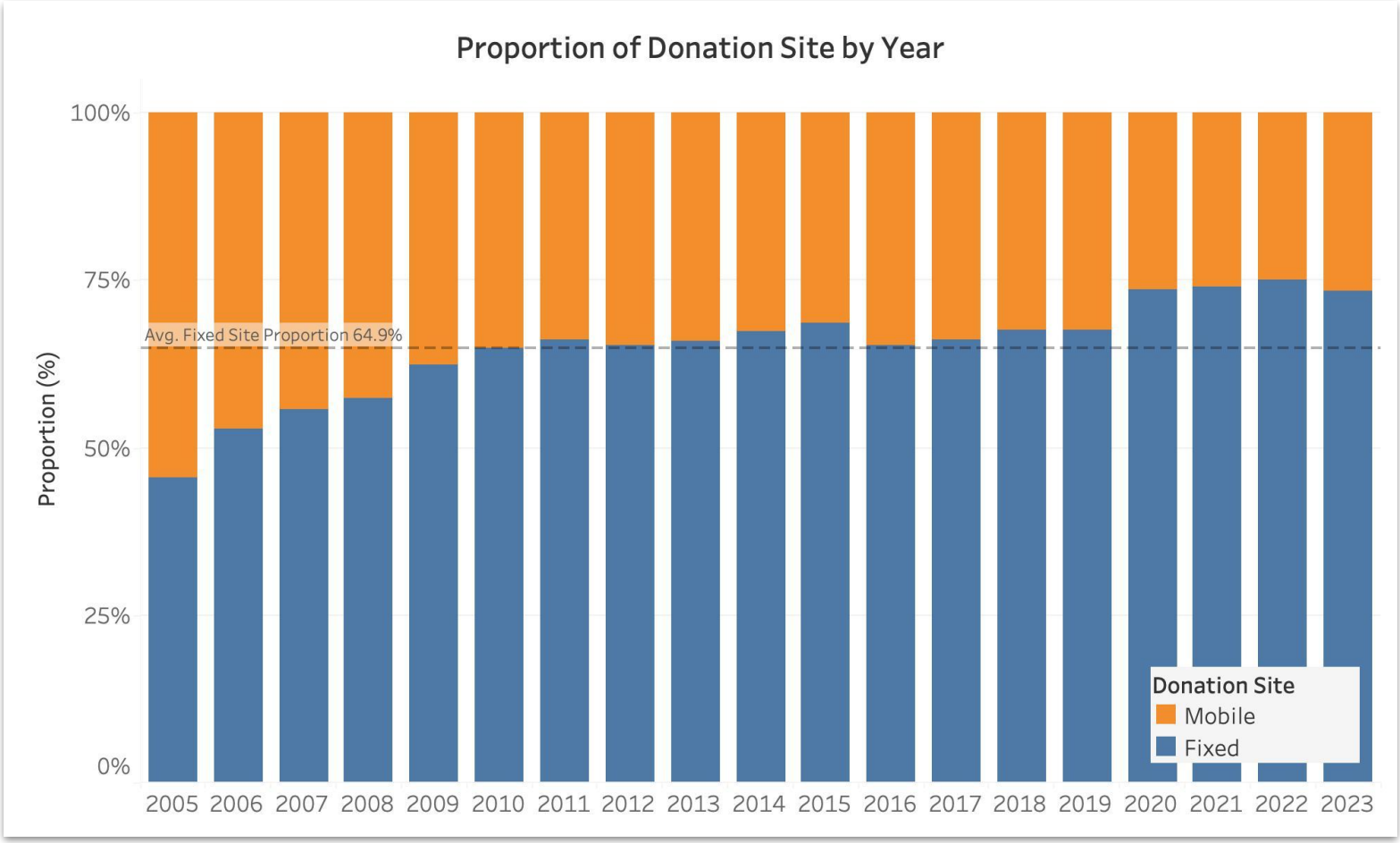
Exploratory Data Analysis (EDA) - by year/site



- The gradual decrease in donations from the youngest two age groups (16-19, 20-29) can be observed with a steep decline in 2020
- Based on the proportion of mobile-site donations, high school and university donations drop significantly in 2020 from **47% to 17%**
- This steep decrease in the proportion of donations from the youngest age groups may be explained by the national lockdown procedures that largely affected schools and universities during the COVID-19 pandemic in 2020

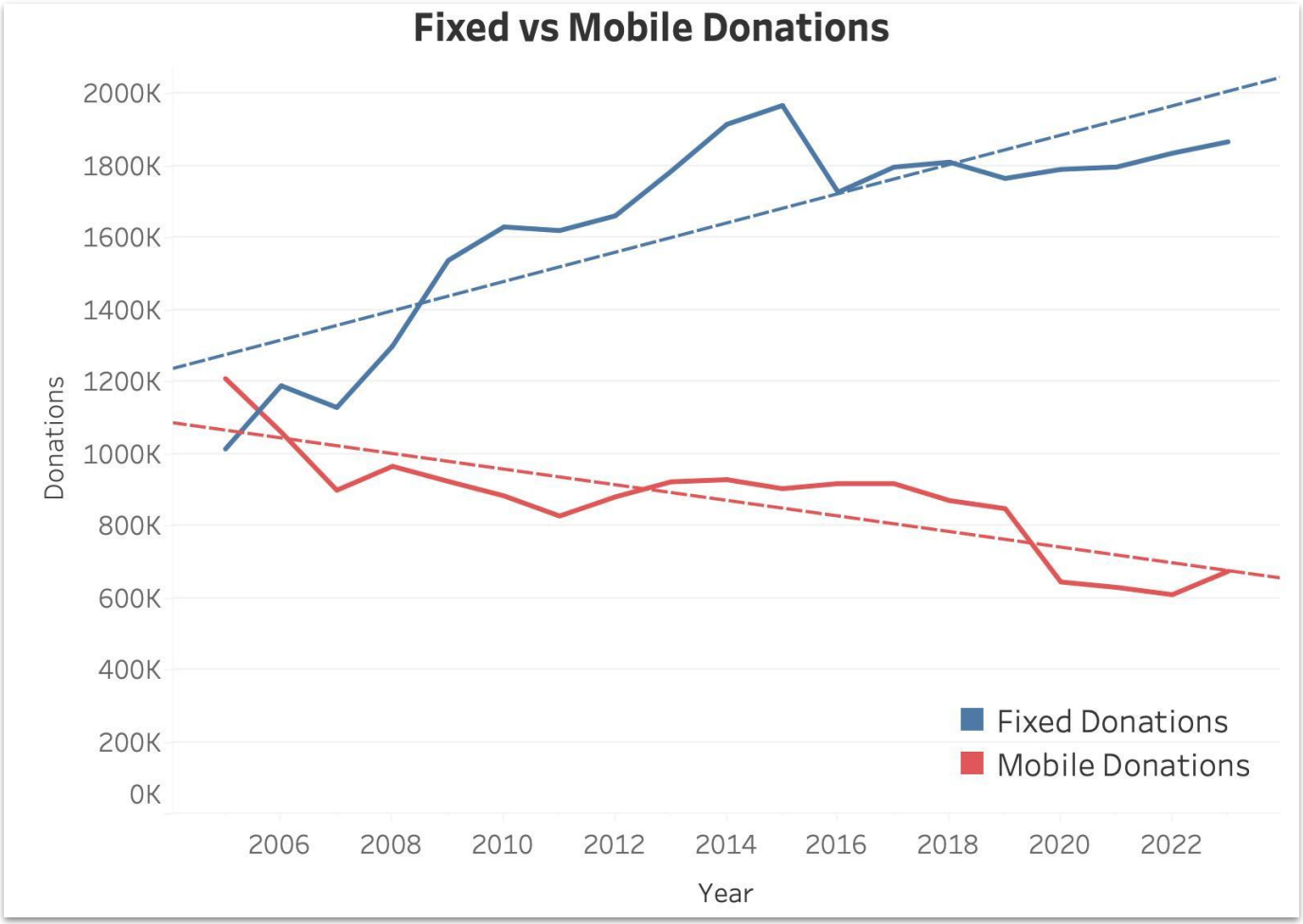
Exploratory Data Analysis (EDA) - by year/site

- The proportion of fixed-site donations averages 64.9%
- This figure increases to reach 75% of total donations in 2022 from the initial minimum of 45% in 2005

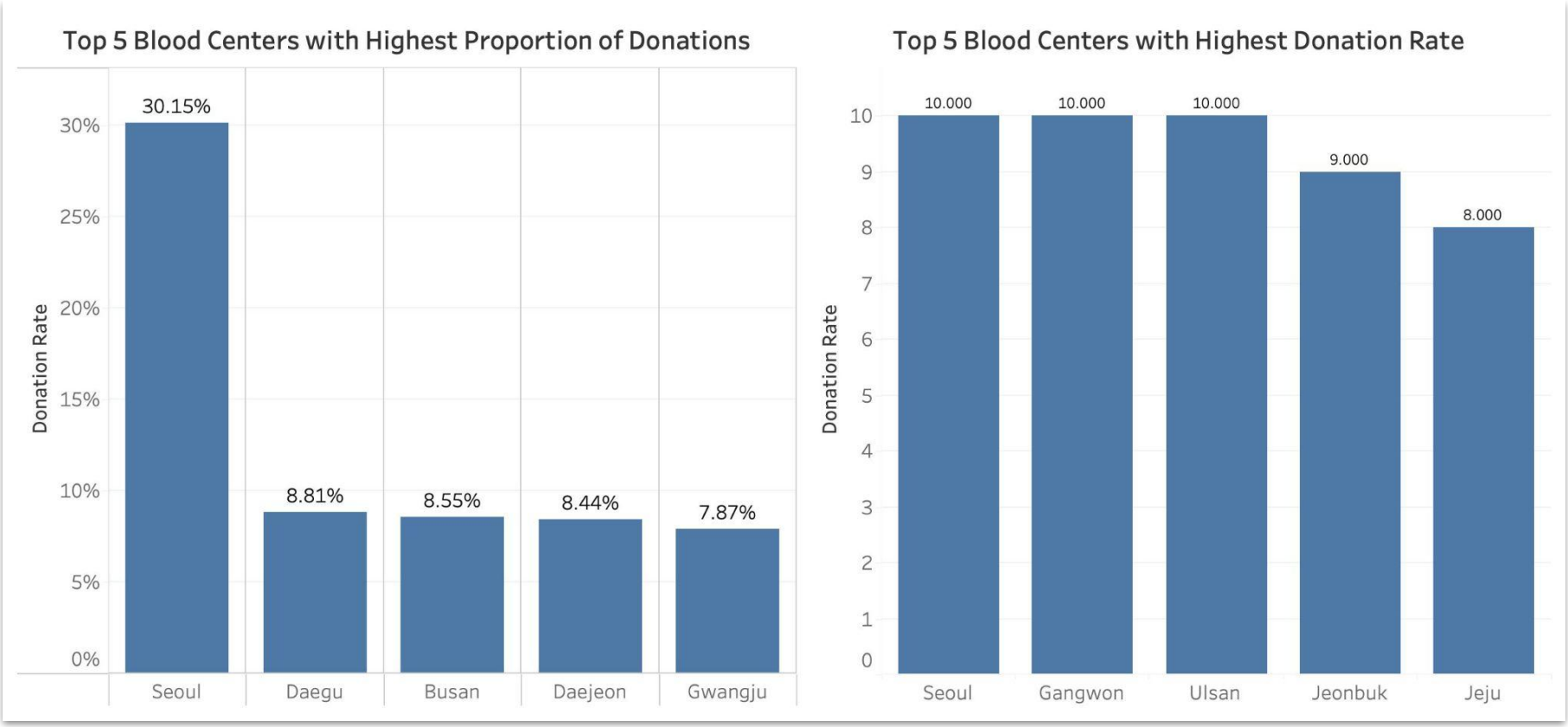


Exploratory Data Analysis (EDA) - by year/site

- In 2005, the donation count starts with higher figures for mobile donations than fixed-site donations
- The count for fixed-site donations immediately overturns and shows an upward trend from 2005 to 2023
- Inversely, the count for mobile donations has a gentle downward trend from 2005 to 2023

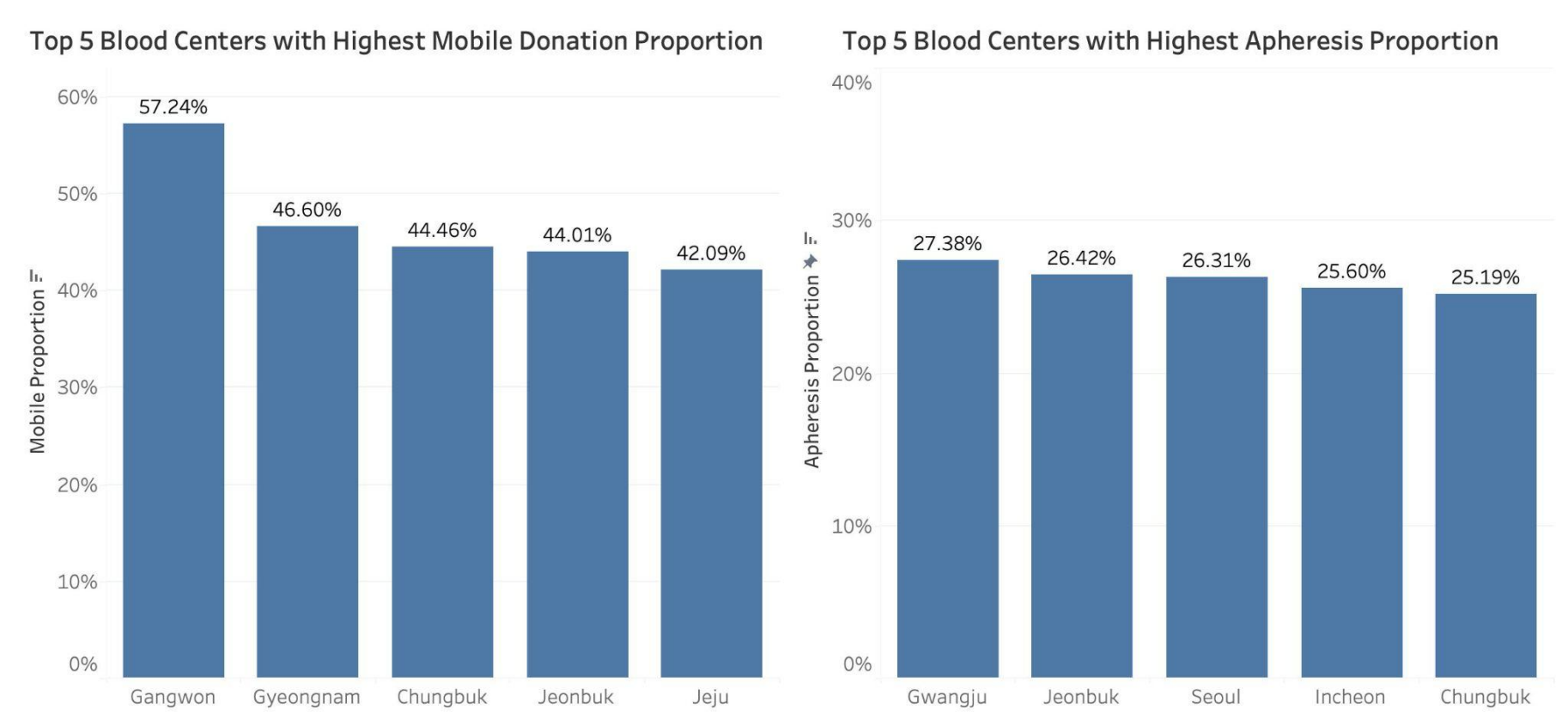


Exploratory Data Analysis (EDA) - by region



- Top 5 regions account for over 65% of total donations
- Seoul, Gangwon, and Ulsan had the highest donation rate among the eligible population at 10%

Exploratory Data Analysis (EDA) - by region



- Gangwon blood center is the only region with higher mobile proportion than fixed-site donation
- Gwangju blood center has the highest apheresis donation proportion with 27% of total donations

paCe: Construct Stage

Prepare Modeling

- The goal of this study is to identify the level of importance for each variable to high donation.
- Build four (4) classification models and select F1 score as the main evaluation score.
- Convert the input features into ratio (considering high variance between features)
- Drop the 'apheresis' donation ratio (to avoid multicollinearity)
- Perform following tests to compare effectiveness of different training methods
    a. Evaluate the effectiveness of 'SMOTE (Synthetic Minority Over-sampling Technique)' in handling imbalanced classes in classification tasks (using the Decision Tree model).
    b. Evaluate differences between 'Grid Search Cross-Validation' or 'Randomized Search Cross-Validation' methods to find the best parameters (using the Random Forest model).

Construct Model

Target (X)

High donations (1, 0) → 25%(1) : 75%(0)

Feature (y)

whole blood, fixed-site, mobile_highschool, mobile_university,
mobile military_camp, others, age groups (under 20, 20s, 30s, 40s, over 50s)

Model

a. Logistic Regression	b. Decision Tree Classifier
c. Random Forest Classifier	d. XGBoosting Classifier

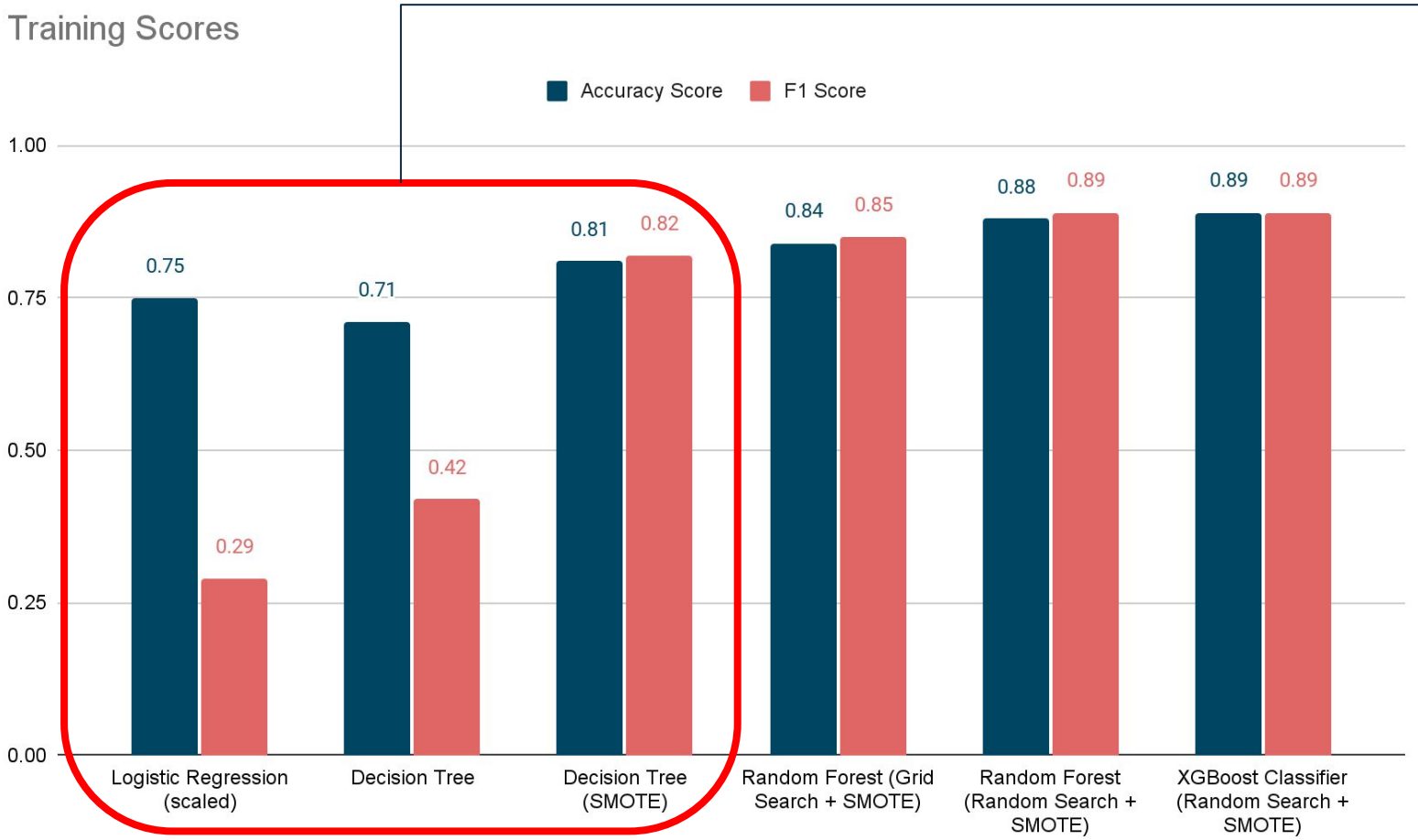
Scaling

Standard Scaler (for logistic regression model only) → SMOTE (Upsampling)

Validation

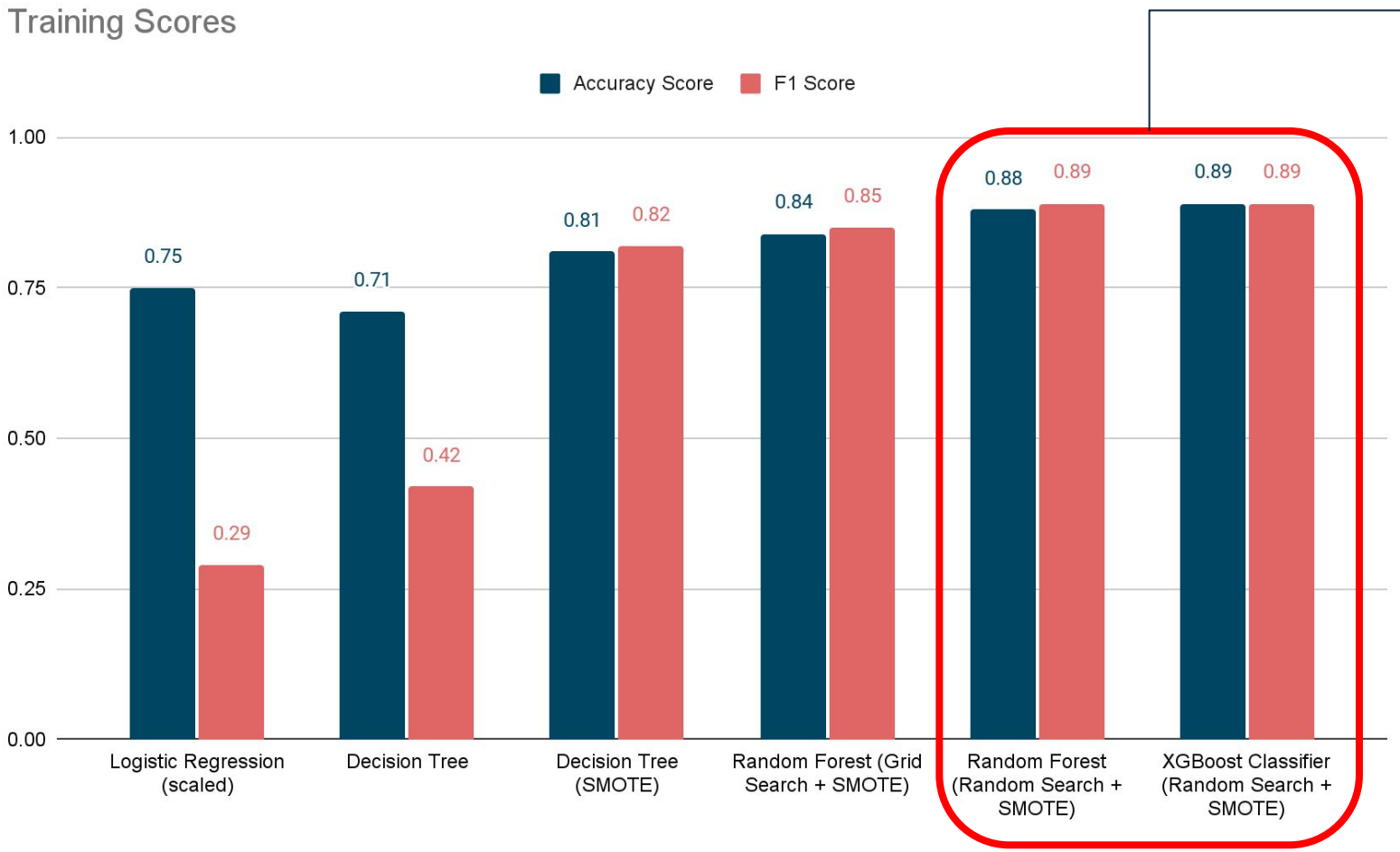
Find best cross validation score using **'StratifiedKFold'** and **'Grid Search'**

Model Results



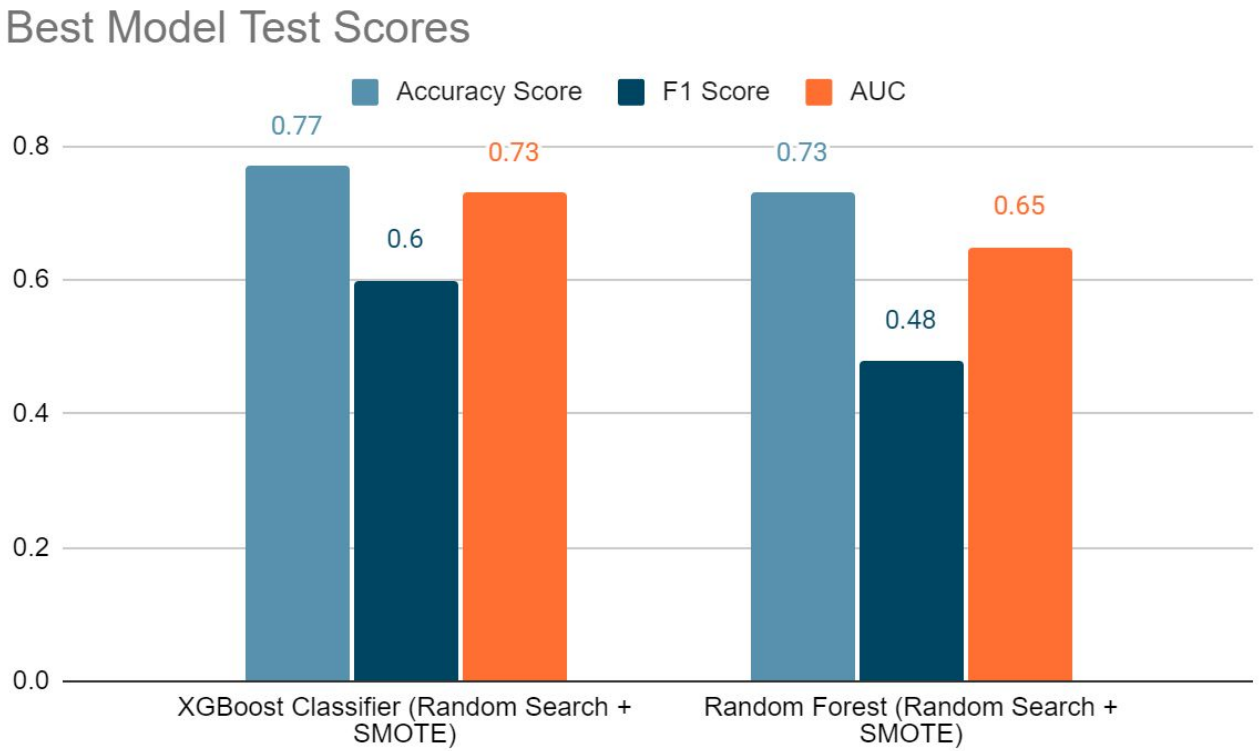
- The Logistic Regression model was unable to explain the data properly (highly due to complexity of the data)
  - high accuracy score and low F1 score indicates the model was unable to capture characteristic of the imbalance data
- SMOTE applied model (Decision Tree + SMOTE) showed fairly balanced performance in 'recall' and 'f1' than non-SMOTE model.

Model Results (continued)

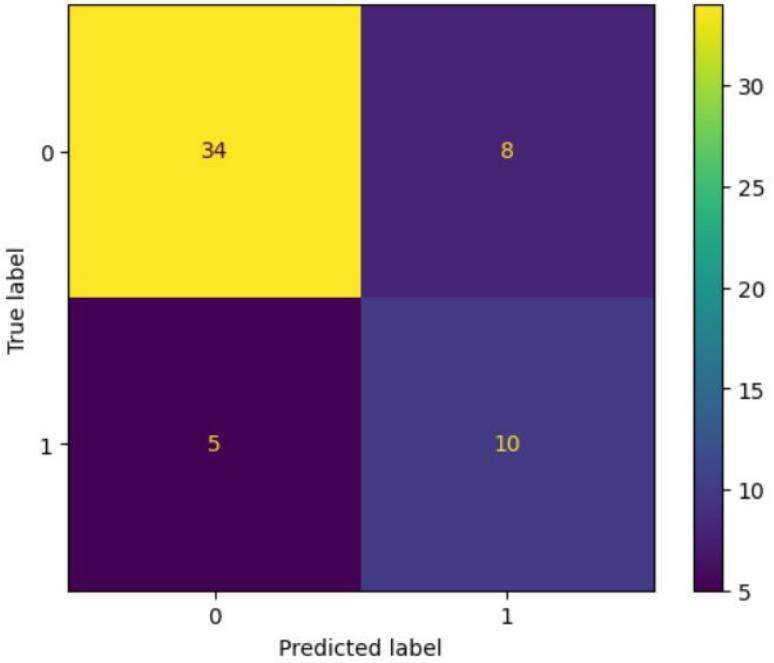


- XGBoost classifier showed the best cross-validation F1 score with less time
- Random forest model using randomized search CV and SMOTE-applied model also showed high score with less cost (time)
  - Random Forest + Grid Search → 8 minutes
  - Random Forest + Randomized Search → less than 2 minutes

Final Model - XGBoost Classifier



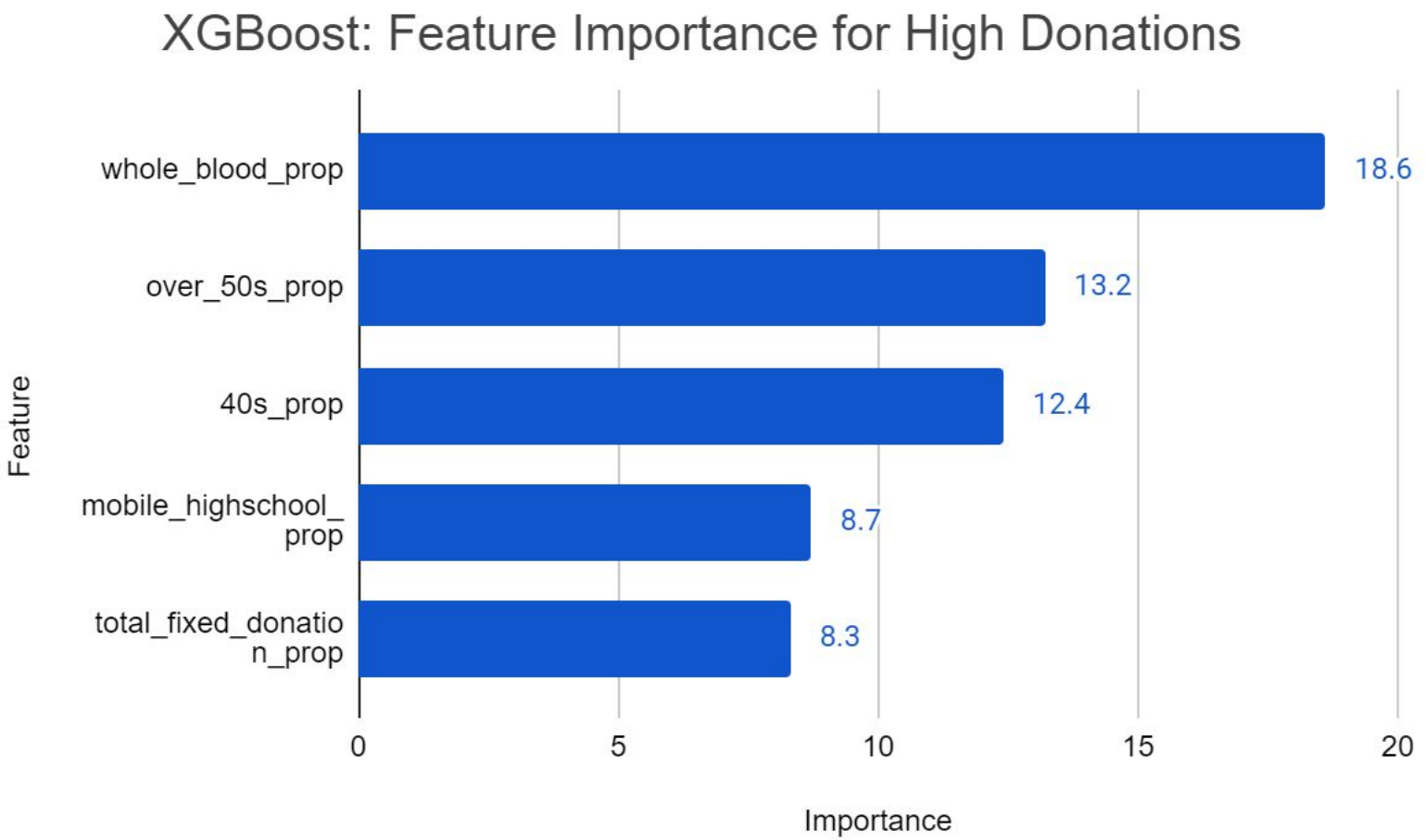
<XGBoost model confusion matrix>



- In the test data, XGBoost classifier model showed higher score with F1 score of 0.6.
- With an AUC score of 0.7381, this model performs better than random guessing 73.81% of the time.
  → Strong ability to classify between the two classes

pacE: Execute Stage

Extract Feature Importance



- In XGBoost model, ‘whole blood proportion’, ‘over 50s proportion’, ‘40s proportion’, ‘high school proportion’, and ‘fixed donation proportion’ have the highest importances.
- These variables are most helpful in predicting ‘high blood donation’ in XGBoost model.

Conclusion (Recommendations)

<Conclusion>

- The XGBoost model achieved an F1-score of 60%, accuracy of 77%, and AUC-score of 73% on the test set.
- **(Limitation)** The model might be overfitted. May need to collect more data to improve model.
- Extracted feature importances from the XGBoost model that are most helpful in predicting high blood donation.
  - Proportion of whole blood donations
  - Age group: Proportion of over 50s and 40s
  - Location: Proportion of 'High school' and 'Fixed-site'

<Recommendations>

- Plan strategies to retain a sufficient number of whole blood donations throughout the year
- To achieve higher performance (donations), consider various campaigns targeting middle-aged donors.
- Plan strategies to improve the proportion of donations at the 'fixed site'. Collecting sufficient donations at the fixed site will contribute to stable performance.
- If the center wants to diversify the mobile donation, interact with local communities to hold blood drives at high school. Delivering the message of blood donation to high school students will help recruit new donors.

Thank You

Reference

Full Project Code: Github
Full EDA Visualization: Tableau

<Contributions>

- **Yoojin (Audrey) Jung** - Data Analyst (Professional)
  - **(Contribution) EDA (processing & visualization), construct model (logistic, decision tree)**
  - (Skills) Data Analytics, Visualization (Tableau), SQL, R, Python
  - (LinkedIn) https://www.linkedin.com/in/yoojin-jung/

- **Hyunjin (Austin) Kang** - Data Scientist (Associate)
  - **(Contribution) Data collection, EDA (feature engineering & visualization), construct model (random forest, XGboost)**
  - (Interest Area) Public Health
  - (Skills) Machine Learning, Python, SQL
  - (LinkedIn) www.linkedin.com/in/austinkang0702