Data analysis report on ‘Predict Blood Donation’ Dataset

Business Problem

Blood transfusion saves lives - from replacing lost blood during major surgery or a serious injury to treating various illnesses and blood disorders. Ensuring that there's enough blood in supply whenever needed is a serious challenge for the health professionals. According to WebMD," about 5 million Americans need a blood transfusion every year". Our dataset is from a mobile blood donation vehicle in Taiwan. The Blood Transfusion Service Center drives to different universities and collects blood as part of a blood drive. We want to predict whether or not a donor will give blood the next time the vehicle comes to campus in March 2007.

Project Report Summary

Overview

The analysis focuses on creating a comprehensive data analysis report to address a specified problem statement. The primary tasks include understanding the dataset structure, preprocessing the data, and mitigating potential risks during the analysis.

Key Findings

1. **Dataset Structure and Cleaning**
    - The dataset consists of columns with 576 values each, with no missing values or null entries.
    - All columns contain integer data types, simplifying preprocessing.
    - An irrelevant column, Unnamed: 0, containing only sequential numbers, was removed to optimize the analysis.
2. **Challenges and Risks**
    - **Feature Scope**: The dataset lacks demographic or behavioral data, which limits the predictive accuracy and generalizability of the analysis.
    - **Biases**: Uneven representation of donors poses a risk of skewed results, potentially impacting model performance and insights.
3. **Next Steps**
    - Conduct advanced feature engineering to improve dataset richness.

- Implement methods to address data imbalance, such as resampling or using model techniques robust to biases.
- Build models and evaluate their performance, ensuring appropriate metrics are used to account for dataset biases.

Exploratory Data Analysis by plotting various graphs

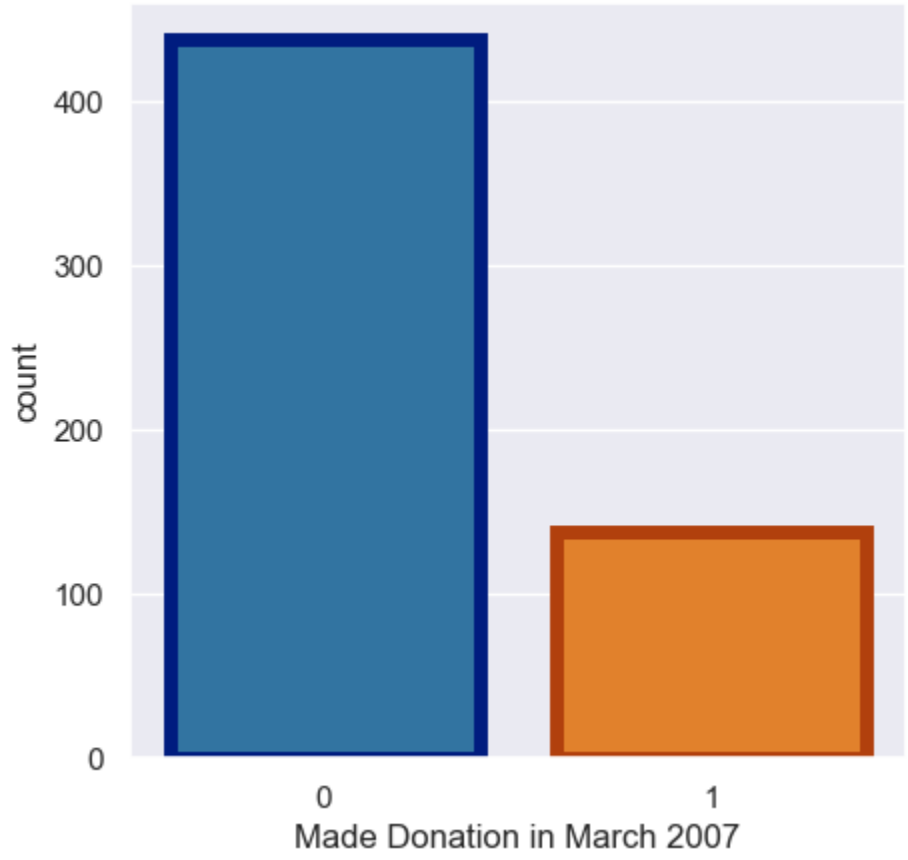
1) Heatmap showing correlation between features

Insights:

- There is a perfect positive correlation (correlation coefficient = 1.00) between the "Number of Donations" and "Total Volume Donated (c.c.)". This is expected because the total volume donated is directly derived from the number of donations. As a result, these variables provide similar information, and one of them could be removed to simplify the model without losing predictive power.

- "Months Since Last Donation" and "Months Since First Donation" show noticeable correlations with whether a donor made a donation in March 2007. This indicates that time-related factors, such as recency and donation history, might be significant predictors of a donor's likelihood to donate in the future.

2) Count Plot for Made donation in March 2007

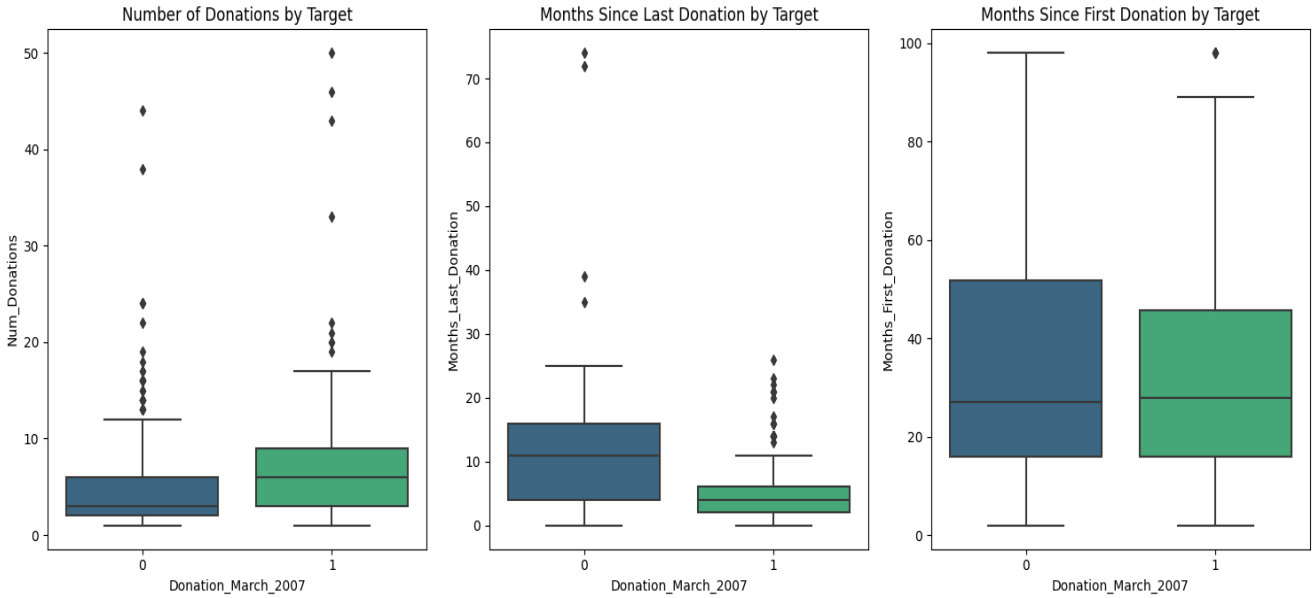


Insights:

- The number of donors who did not donate in March 2007 (class 0) is significantly higher than those who did (class 1). This indicates a class imbalance in the dataset, which could affect model performance. Addressing this imbalance with techniques like oversampling the minority class or under sampling the majority class might be necessary for better predictive accuracy.

- A relatively small number of donors made donations in March 2007. This suggests that predicting future donations may require a focus on identifying the characteristics of this smaller group to build an effective classification model. Variables like recency of last donation and overall donation history may play a key role in distinguishing these donors.

3) Boxplots for feature-target relationships

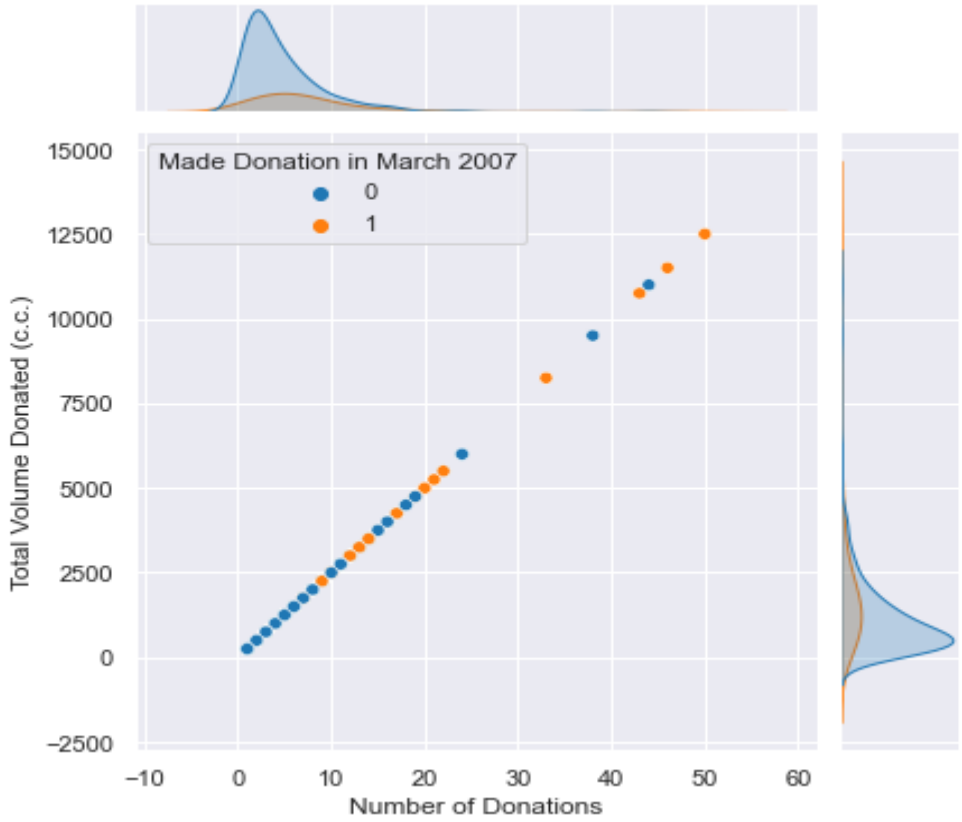


Insights:

- Donors who made a donation in March 2007 (Donation_March_2007 = 1) tend to have a slightly higher median number of donations compared to those who did not donate.

- Donors who did not donate in March 2007 (Donation_March_2007 = 0) generally have a larger gap (higher months) since their last donation, indicating potential donor disengagement.

4) Join plot between Number of Donations and Total Volume Donated

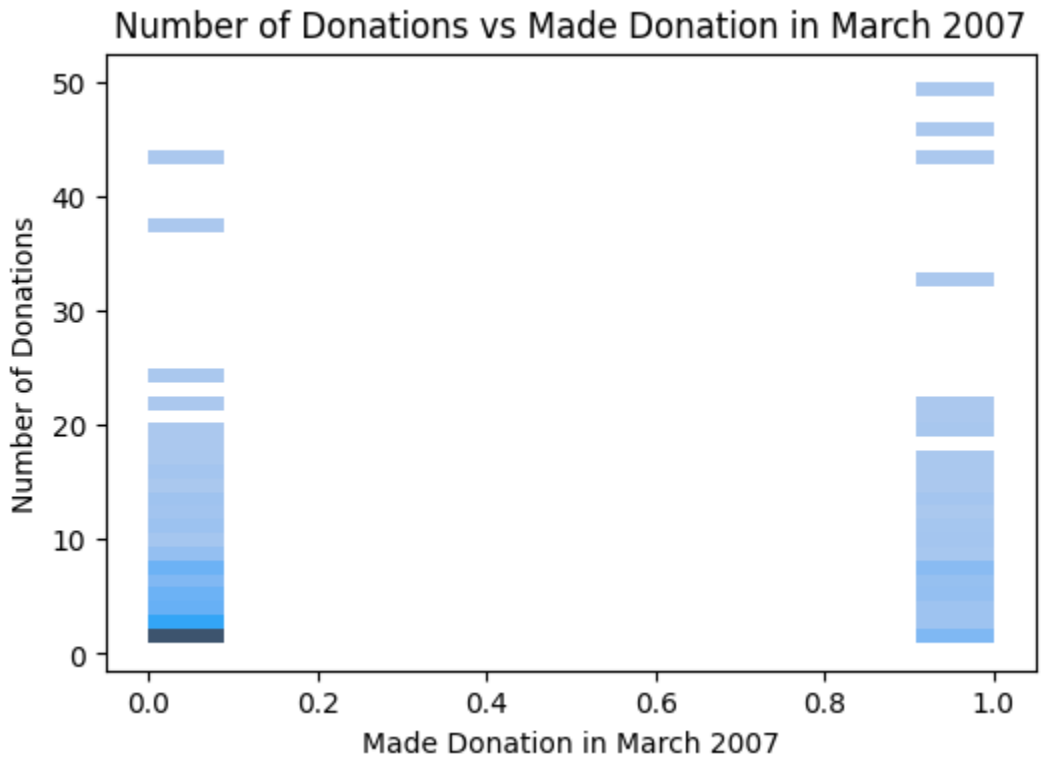


Insights:

- The join plot shows that Months_First_Donation and Months_Last_Donation are positively correlated, indicating that donors who have been active longer are also more likely to have donated recently.

- The hue for Donation_March_2007 highlights that donors who made a donation in March 2007 (Donation_March_2007 = 1) tend to cluster in specific ranges for features like Num_Donations, suggesting a pattern of engagement tied to their donation history.

5) Strip plot or a variation of a categorical scatter plot for Number of Donations vs Made Donation in March 2007

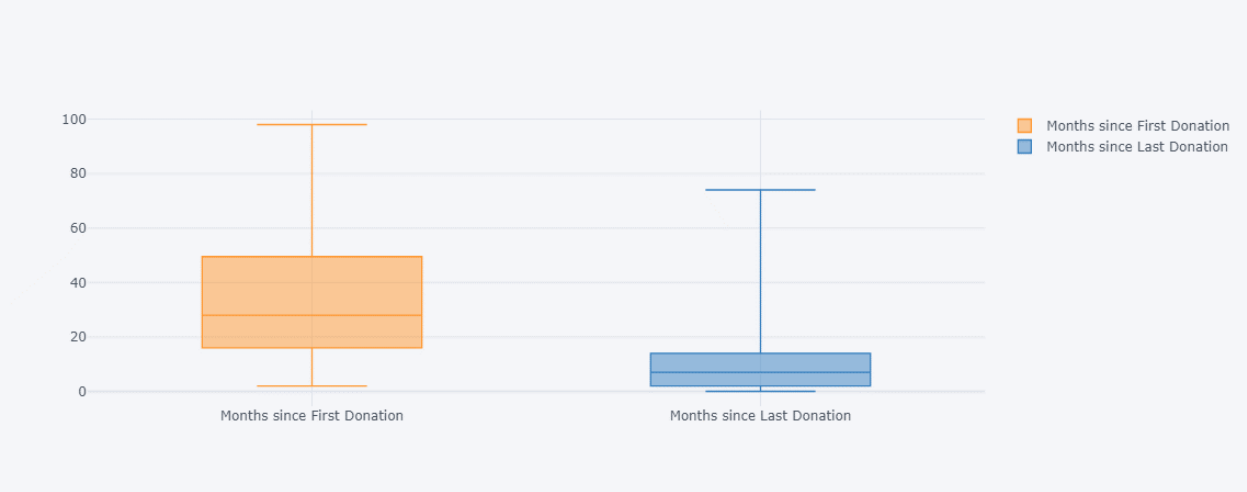


Insights:

- Donors who made a donation in March 2007 generally had a higher number of past donations compared to those who did not.

- There is significant overlap in donation history between the two groups, suggesting additional factors may influence donation likelihood.

6) Box Plot comparison between first and last donation

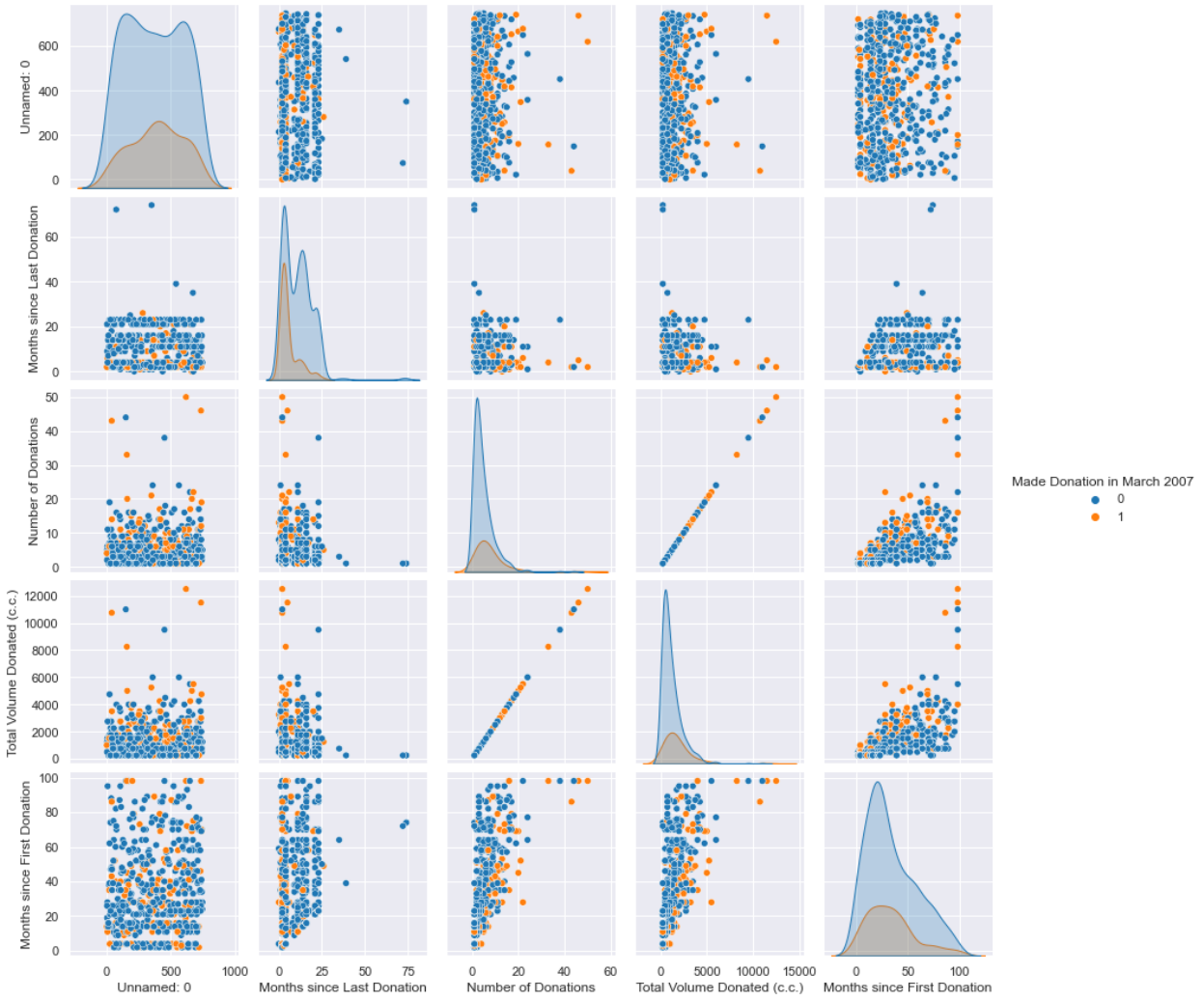


Insights:

- The "Months since First Donation" boxplot shows a wide range, with values spanning from very recent donations to donors whose first donations were nearly 100 months ago. This indicates that the donor pool has a mix of relatively new and long-standing donors, suggesting varying levels of engagement with the blood donation service.

- The "Months since Last Donation" boxplot shows a much narrower range compared to the "Months since First Donation." This suggests that most donors have given blood recently, with fewer outliers who haven't donated in a long time. This could indicate that a significant portion of donors are frequent or semi-regular contributors, which is promising for predicting future donations.

7) Pair Plot

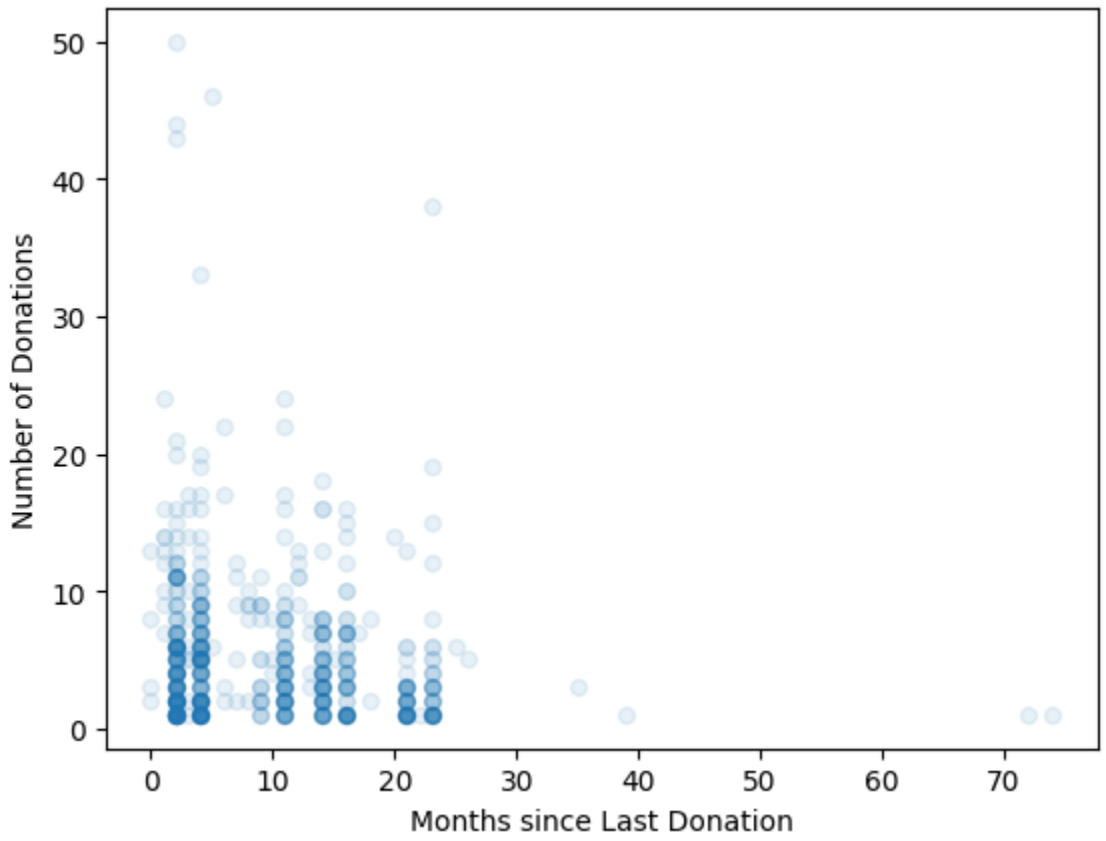


Insights:

- Donors who have a higher "Number of Donations" (visible along the diagonal and in scatterplots) are more likely to donate again in March 2007 (orange points). This suggests that past donation frequency is a strong predictor of future donation likelihood.

- In the "Months since Last Donation" column, donors represented by orange points (those who donated in March 2007) tend to have lower values compared to blue points. This indicates that recent donors are more likely to donate again, highlighting the importance of targeting individuals who have donated recently for future campaigns.

8) Scatter plot for Number of Donations vs Months since Last Donation

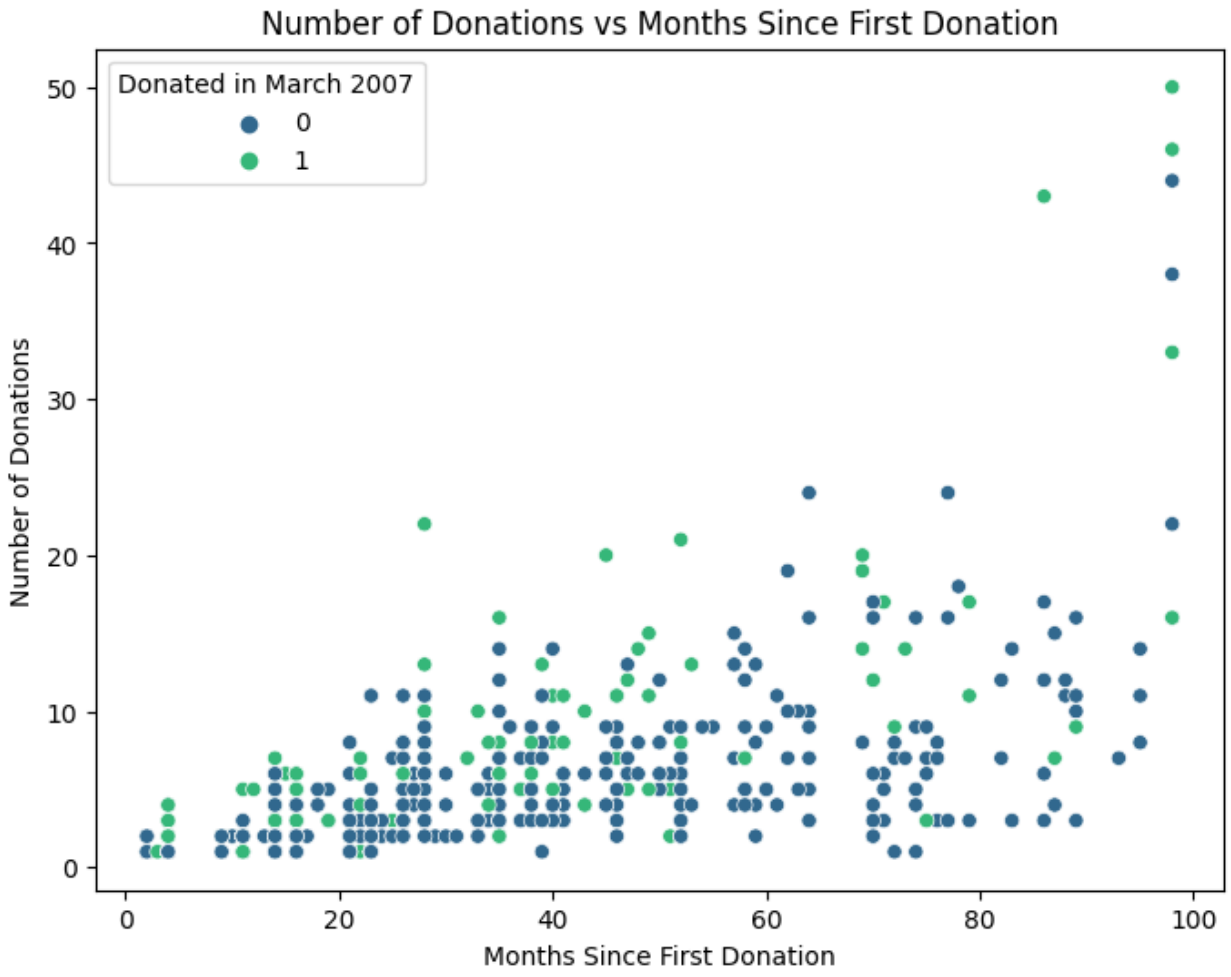


Insights:

- Donors with a higher "Number of Donations" tend to have a shorter "Months since Last Donation." This indicates that more frequent donors are likely to donate more consistently and recently.

- As the "Months since Last Donation" increases, the "Number of Donations" decreases significantly, with very few donors contributing after long gaps. This suggests that re-engaging donors who haven't donated recently could be a challenge but might have potential for targeted campaigns.

9) Scatter plot for Number of Donations vs Months since First Donation



Insights:

- Donors who have been donating blood for a longer time (higher months since the first donation) generally show a trend of more donations overall. This suggests that loyal, long-term donors contribute significantly to the blood supply.

- A significant portion of donations comes from individuals who have only recently started donating (within the first 20 months). This implies that a substantial effort is required to retain these new donors to sustain or increase blood supply levels over time.

Conclusion

The analysis of the "Predict Blood Donation" dataset revealed significant insights into donor behavior, emphasizing the challenges in predicting future blood donations. Key findings demonstrate strong correlations between the "Number of Donations" and "Total Volume Donated," as well as time-based features like "Months Since Last Donation" and "Months Since First Donation," which are crucial predictors for future donation likelihood. Class imbalance in the dataset, with fewer donors in the positive class (those who donated in March 2007), presents a major challenge, necessitating advanced balancing techniques like oversampling. Exploratory analyses using visualizations such as scatter plots, heatmaps, boxplots, highlighted that frequent, recent donors are more likely to donate again, while long gaps between donations often correlate with lower engagement. While the dataset is clean and structured, its limited feature diversity restricts the generalizability of the models. Therefore, incorporating demographic, behavioral, or motivational factors in future analyses would enhance prediction accuracy and offer deeper insights into donor retention strategies. Ultimately, this analysis underlines the importance of targeted campaigns focusing on recent and frequent donors to ensure a sustainable blood supply and improve healthcare outcomes.