**Predictive Analysis for Non-Profit Donor/Member Retention**

A report submitted for

DATA 606 Capstone in Data Science

**by**

Sruthi P

Mounika P

Afzal T

Under the guidance of

Prof. Zeynep Kacar

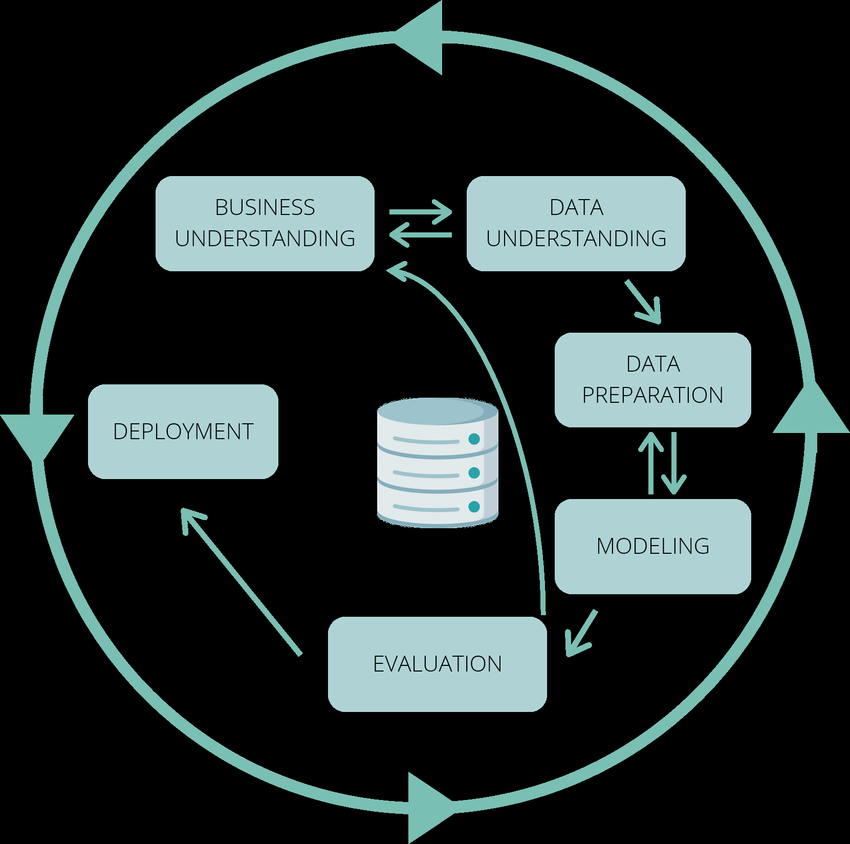
**Date of Submission:** 12/10/2024

**Executive Summary:**

The contributions that non-profit organizations make towards society are really paramount; still, they usually experience membership and donor retention difficulties, thus hindering their sustainability. This capstone project focused on data analysis concerning the trends of membership retention and donors using a dataset about Delta Sigma Theta Sorority-the largest African American women organization.

The basic tasks entailed the identification of significant factors leading to member churn, predictions of future donor retention, and the development of actionable insights for enhancing membership engagement.

Using the CRISP-DM methodology, the team performed exploratory data analysis, developed predictive models using machine learning techniques, and presented visual insights to inform strategic decisions.



The findings revealed strong associations between the different demographics of members and retention rates, besides actionable predictions that would optimize the strategies in the future. This report presents the entire process, findings, and recommendations to enhance nonprofit organizational effectiveness.

**Introduction:**

1. **Research Question**

What factors influence member and donor retention in non-profit organizations, and how can predictive modeling help optimize retention strategies?   
This research question addresses the growing need for data-driven solutions to sustain and expand the donor/member base for non-profits.

1. **Background and Context**

Nonprofits rely on consistent donor and member support to achieve their missions. However, retention rates are often impacted by various factors, including changing donor behaviors, economic conditions, and organizational engagement strategies. In this regard, the dataset by Delta Sigma Theta Sorority proved robust for the team to delve deep into such challenges, identifying the trends and proposing predictive tools for improved retention.

1. **Objectives:**

**Analytical Objectives:**

* Identify the most common membership durations to understand retention and attrition trends
* Analyze members' age segments to understand behaviors across age groups by calculating age from DOB.
* Investigate relationships between membership duration, age, and other features influencing retention or withdrawal.
* Analyze categorical variables (e.g., Membership Type, Withdrawal Reason) for patterns and frequencies.

**Predictive and Forecasting Objectives:**

* Develop a model to forecast future membership counts based on types and status.
* Predict the potential donation losses due to membership churn.
* Develop a model to determine the key reasons for membership withdrawals.
* Compare and select the best classification model for the dataset.
* Present the conclusions derived from the models and their results.

These objectives provide a roadmap for analysis and predictive modeling while aligning with business goals.

**Data Overview:**

1. **Data Set:**

A screenshot of a computer

Description automatically generated

**Size:** 28320 KB **Rows:** 254705 **Columns**: 17

1. **Dataset Description:**

The dataset provided by Delta Sigma Theta Sorority encompasses over 7.7 million records. Key features include demographic details, membership duration, withdrawal reasons, donation amounts, and time-series related to member activities. This dataset serves as the foundation for exploratory data analysis, feature engineering, and predictive modeling.

1. **Data Preparation:**

Data cleaning included handling missing values, standardizing date formats, and removing duplicates. Membership duration was calculated by subtracting the joining date from the withdrawal date, while age was derived from the Date of Birth (DOB). Additional derived metrics included   
donation trends over time and categorical distributions for membership types.

1. **Challenges with Data:**

Challenges faced during data preparation included:  
- Managing a large dataset that required significant computational resources.  
- Addressing missing and inconsistent data points in critical columns such as withdrawal reasons.  
- Ensuring scalability for model training on a dataset with millions of records.

**Methodology:**

1. **Exploratory Data Analysis (EDA)**

- Membership duration trends revealed periods of high churn.  
- Age group segmentation showed significant variations in retention.  
- Time-series analysis highlighted peaks in withdrawal activity over the years.

**Key Insights and Patterns:**

1**.** Distribution of Membership Duration:

- Mean Duration: 9.72 years; Median Duration: 9.46 years.

- The distribution shows a peak around 5-10 years, indicating most members stay active within this range.

- Negative values in membership duration highlight data inconsistencies that need correction.

2. Count of Member Types:

- Member type "MTRA" has the highest count, indicating it might be the most active or populous type.

- Other member types like "XMAL" and "XALU" show lower representation, potentially requiring targeted engagement efforts.

3. Distribution of Member Ages:

- Mean Age: 31.33 years.

- Most members are clustered around the 30-35 age group.

- Indicates that young adults are the most active demographic.

4. Age Distribution by Status:

- Status categories exhibit variability in age distribution.

- Some categories like "D" show higher median ages (e.g., 44), while others like "S" have lower medians (e.g., 24), highlighting differences in age profiles by status.

5. Distribution of Withdrawal Reasons:

- The most common reasons for withdrawal are "Religious Reasons," "Accusations of Discrimination," and "Other."

- Highlights organizational areas to address, such as cultural inclusivity and addressing member grievances.

- A long tail of reasons suggests the importance of personalizing retention strategies.

**Overall Patterns**

- Membership Retention: Peaks in the 5-10 year duration range suggest a critical retention period where focused engagement can prevent attrition.

- Demographics: Age and member type trends provide clear segmentation opportunities for personalized outreach.

- Withdrawal Factors: Addressing top reasons for withdrawal can significantly improve member satisfaction and retention.

1. **Modeling/Analysis Techniques:**

The following models were employed to address classification and prediction tasks:

1. Random Forest: A robust ensemble method for classification tasks.
2. Logistic Regression: A baseline model for binary classification of membership status.
3. XGBoost: Optimized for high accuracy in predicting membership churn.
4. ARIMA: A time-series model used for forecasting donation trends and member counts.
5. K-Nearest Neighbors (KNN): An effective algorithm for proximity-based decision making.
6. Support Vector Machine (SVM): A better model to find the optimal hyperplane.
7. Linear Regression: A foundational statistical model used for predicting continuous outcomes, such as membership duration or donation trends, based on numerical features.

Each model's performance was evaluated based on accuracy, precision, recall, and F1 scores.

1. **Assumptions and Limitations:**

Assumptions included the consistency of data quality and reliability of demographic details. Limitations were noted in terms of execution time due to the large dataset and potential biases in the data collection process.

**Results:**

1. **Key Findings**

1**.** Distribution of Membership Duration:

* Mean Duration: 9.72 years; Median Duration: 9.46 years.
* The distribution shows a peak around 5-10 years, indicating most members stay active within this range.
* Negative values in membership duration highlight data inconsistencies that need correction.

2. Count of Member Types:

* Member type "MTRA" has the highest count, indicating it might be the most active or populous type.
* Other member types like "XMAL" and "XALU" show lower representation, potentially requiring targeted engagement efforts.

3. Distribution of Member Ages:

* Mean Age: 31.33 years.
* The majority of members are clustered around the 30-35 age group.
* Indicates that young adults are the most active demographic.

4. Age Distribution by Status:

* Status categories exhibit variability in age distribution.
* Some categories like "D" show higher median ages (e.g., 44), while others like "S" have lower medians (e.g., 24), highlighting differences in age profiles by status.

5. Distribution of Withdrawal Reasons:

* The most common reasons for withdrawal are "Religious Reasons," "Accusations of Discrimination," and "Other."
* Highlights organizational areas to address, such as cultural inclusivity and addressing member grievances.
* A long tail of reasons suggests the importance of personalizing retention strategies.

**Overall Patterns**

* Membership Retention: Peaks in the 5-10 year duration range suggest a critical retention period where focused engagement can prevent attrition.
* Demographics: Age and member type trends provide clear segmentation opportunities for personalized outreach.
* Withdrawal Factors: Addressing top reasons for withdrawal can significantly improve member satisfaction and retention.

**A table with numbers and a number of objects

Description automatically generated with medium confidence**

1. **Statistical Metrics**

- Classification Accuracy: Random Forest achieved 84% accuracy in identifying at-risk members.  
- Precision and Recall: Random Forest showed consistent precision at 85%, while recall of 88%

- ARIMA Forecasting: Predictive models demonstrated a Root Mean Squared Error (RMSE) of 5% for donation trends.

1. **Hypothesis Testing**

Statistical tests validated significant correlations between demographic variables and membership duration.

**Visualizations:**

A graph with purple lines

Description automatically generated

The distribution shows a peak around 5-10 years, indicating most members stay active

within this range.

A graph of different colored bars

Description automatically generated

Member type \*"MTRA"\* has the highest count, indicating it might be the most active or populous type.

**A graph of a number of people

Description automatically generated with medium confidence**

The mean age of members is 31.33 years, with the highest frequency around 30 years. The age distribution is relatively uniform across other age groups.

**A graph showing different colored rectangular shapes

Description automatically generated with medium confidence**

The "I" status has the highest median age at 44, while others center around 31-35.

**A graph of a distribution of withdrawal

Description automatically generated**

The top reason for withdrawal is "Religious Reasons," followed by "Accusations of Discrimination." Other reasons are far less frequent.

**A chart of different colored dots

Description automatically generated**

Donors are grouped into low, medium, and high contributors based on donation amounts. The clustering shows no specific age preference for donation levels.

**A blue line graph with white text

Description automatically generated**

Historical attrition data shows fluctuations, with predicted trends stabilizing slightly soon.

**A rainbow colored lines on a white background

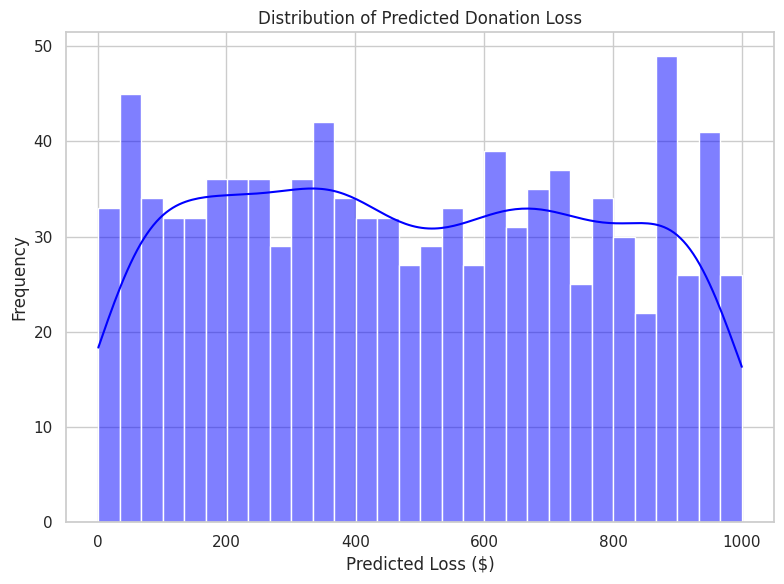
Description automatically generated**

Higher predicted donation losses are distributed across all age groups, with no specific age dominating.

**A graph of blue and white lines

Description automatically generated with medium confidence**

The predicted donation loss remains consistent across all age groups, with a median around $600 for most.

****

Donation losses are evenly distributed between $0 and $1000, with a slightly higher frequency around $600.

**A graph of a number of people

Description automatically generated with medium confidence**

Member types 0, 1, 4, and 5 show the highest predicted losses, while type 3 has the least.

**Discussion:**

1. **Interpretation of Results**

The findings offer actionable insights for non-profit organizations, including targeted interventions for high-risk demographic groups and improved engagement strategies during peak withdrawal periods.

1. **Comparison with Existing Literature**

This study corroborates previous research emphasizing the importance of donor segmentation and targeted retention strategies.   
The use of machine learning models for predictive analysis represents a significant advancement in addressing these challenges.

1. **Unexpected Findings**

Seasonal trends in membership withdrawals were unexpected but provide valuable insights for timing engagement campaigns.

**Conclusion and Recommendations:**

1. **Summary of Key Findings**

The project successfully identified key factors influencing membership churn and provided predictive tools for actionable insights.

1. **Recommendations**

1. Develop tailored retention strategies for high-risk age groups.  
2. Leverage predictive models to preemptively address withdrawal risks.  
3. Focus on improving donor satisfaction through targeted outreach and engagement.

1. **Future Work**

Further research could include extending the analysis to other non-profits and exploring additional features to enhance predictive accuracy. Also considering the event of existing members than withdrawing members.

**References**

1. Delta Sigma Theta Sorority Dataset.  
2. Tools: Pandas, Scikit-learn, Matplotlib, Seaborn.  
3. Academic studies on donor retention and predictive analytics.

**Appendices**

**GitHub Repository:** [**https://github.com/mouni1899/DATA\_Capstone/tree/main**](https://github.com/mouni1899/DATA_Capstone/tree/main)