Department of Statistics 2024/2025

Capstone Project: Data-driven Approaches for Legacy Analysis and Prospection

Candidate Number: 40781, 38837, 42507

Submitted for the Master of Science,
London School of Economics, University of London

Contents

Executive summary

This study developed various data-driven analysis approaches to alumni legacy donation of LSE and provided feasible suggestions to the PAGE department to promote legacy donation. Based on the available data, the investigation was divided into two parts. In the first part, the analysis was conducted in the scope of all alumni who were present in the legacy pipeline (at least one stages of Enquirer, Intender, Pledge, and Proposal), in which the characteristics of legacy donors were extracted and compared with that of alumni who have joined in the legacy pipeline but have not proposed to donate. This can assist in identifying **potential donors** from those alumni who are already interested in legacy donation. In the second part, the entire alumni population, which has been actively connecting with the PAGE department, was investigated, in order to distinguish **potential participants** of the legacy pipeline. Therefore, the department might target on this group of alumni for promoting the pipeline. To address above research questions, three different approaches were adopted to generate thorough insights into the legacy pipeline and prospects: factor analysis (unsupervised), cluster analysis (unsupervised), and predictive models (supervised), which are the methods suggested in Capstone Project Proposal.

Throughout the project, the main challenge was found to be the quality of datasets provided, in which around 31.2% original data was missing or non-informative and majority of variables were categorical. Therefore, a series of data cleaning and feature engineering was conducted in the first place to build two applicable datasets for the analysis of two research questions. Missing values concentrated in the group of alumni with higher ages or who have passed away. This was likely due to the improvement in the scope of data collection by the PAGE, in which earlier alumni did not have chance to provide more complete information. Moreover, an investigation on the conversion between pipeline stages suggests that only a small portion of alumni have participated in all four stages of the pipeline, and the conversion rates from one of the three early stages to the Proposal stage were 13.58% for Enquirer, 52.03% for Intender, and 70.16% for Pledge. This demonstrates the space for improvement in effectiveness of the legacy pipeline.

The three types of approaches applied in this study provided insights in different aspects. Factor analysis was used to discover the latent structures hidden in datasets, identifying key features contributing to major variations in data. Factor structures were proposed by exploratory factor analysis and then verified by confirmatory factor analysis. Two three-factor models were generated to capture the underlying structures of two sets of data, in which **Enquiry Engagement**, **Alumni Relation**, and **Personal Profile** were factors affecting **potential donors**, and **potential participants** mainly varied in aspects of **Degree**, **Alumni Relation**, and **Personal Profile**. Moreover, cluster analysis successfully categorised populations into groups with distinguishing characteristics and identified potential donors and participants from alumni populations based on these core features. In **potential donors**, cluster analysis used hierarchical clustering as the best model, dividing the population into 2 clusters, while in **potential participants**, it chose K-means clustering and also generated 2 clusters. Last but not least, predictive models performed binary classifications via various methods, including Random Forest, GBDT, XGBoost and Logistic Regression. In the anal-

ysis of **potential donors**, the Random Forest tuned via Random Search was selected as the single best model with the accuracy of 86.25%. In the analysis of **potential participants**, XGBoost was more suitable for this task with the accuracy of 95.27%.

The results obtained from different methods are strongly connected and can bring new insights into the legacy pipeline. In both factor analysis and predictive models, importance variables leading to variance in data were selected. Although some of these features obtained from the two methods were different, they still implied the variables that the PAGE might want to pay more attention to in later data collection. Furthermore, both cluster analysis and predictive models identified a list of potential legacy donors and pipeline participants, and discovered their core features from different technical perspectives. The group division from two models were mostly aligned, but still had a discrepancy. It would be reliable to combine the positive potential donor samples of the two to capture every high-value alumni. The core features from cluster analysis were more about demographic features such as age and marriage, which could easily and quickly classify the potential donors with alumni's complete personal information. Cluster analysis described the profile of the different groups of alumni so as to make customized guidance more precisely. Predictive models focused on the features like recent donating participation behavior and the enthusiasm level of interaction, making an assessment of the potential for alumni to donate legacy. The combination of two methods could enhance the long-term cultivation and short-term conversion of alumni, which enabled the encouragement of the core group with efficient resource allocation.

Overall, this report presents a prudent pipeline for data-driven analysis on legacy donation using factor analysis, cluster analysis, and predictive models, which are replicable and sustainable for future uses. Apart from technical results obtained these models, practical suggestions have been provided to the PAGE department including target populations for promoting the legacy pipeline and improvements in data collection.

1 Introduction

1.1 Objectives and Research Questions

The Philanthropy and Global Engagement division (PAGE) of London School of Economics and Political Science (LSE) is responsible for promoting the legacy donation programme to alumni and current and past staff of the School. Legacy has great significance beyond financial support, since it enhances alumni engagement and demonstrates the contiuity of culture and spirits of the School. This study aims to investigate the key characteristics of existing donors and participants at each stage of the legacy pipeline and propose data-driven approaches to identify the best legacy prospects based on available alumni data. Anonymised real data of active alumni and alumni who has entered the legacy pipeline are provided by the PAGE. The legacy pipeline has four stages in the order of Enquirer, Intender, Pledge, and Proposal, but it is possible to enter at any stage or jump a step. People who have participated the pipeline at any stage may also join the Circle which helps the PAGE to manage their relationship with donors and potential donors.

This study primarily concentrates on two research questions: how legacy donors are different from pipeline participants who have not decided to donate, and how pipeline participants (no matter donated or not) can be distinguished from the alumni population. These questions are addressed by a range of statistical and machine learning techniques including predictive models, factor analysis, and cluster analysis. Methods proposed are assessed in terms of both reasonability and interpretability under the practical context, in which whether an approach can be effectively adopted by the PAGE to improve the legacy marketing programme is the primary criterion for evaluation.

Based on the real alumni data, the current LSE alumni profile composition is analysed and potential legacy participants are identified according to proposed methodologies, so recommendations on the marketing programme are provided to better target on the prospects. To enhance future legacy promotion and alumni connection, suggestions on data collection and maintenance, including useful dimensions and data integrity improvement, are propounded for the PAGE under prudent ethics considerations.

1.2 Summary of Main Results

For **potential donors** (within the legacy pipeline), factor analysis revealed several latent dimensions underpinning donation behaviour, notably age/tenure with the School, degree type, and prior engagement indicators. These factors explained a substantial portion of variance and yielded interpretable profiles such as “Mature, Highly Engaged Alumni” versus “Younger, Low-Engagement Alumni.”

Cluster analysis further segmented pipeline participants into behaviourally distinct groups. High-intent donor clusters typically combined advanced age, prior pledges, and strong event participation, whereas low-intent clusters were younger and less engaged. This segmentation highlighted cohorts such as “High-Value Potential Donors” and “Occasional Participants,” providing a basis for tailored cultivation strategies.

Predictive modelling identified the Random Search–tuned Random Forest as the best-

performing model for donor prediction. Key predictors were age, circle engagement, and participation in other legacy pipeline stages, reinforcing the importance of both demographic maturity and multi-stage involvement for donation likelihood.

For **potential participants** (within the active alumni population), factor analysis identified underlying engagement and affinity dimensions, including relationship manager contact, subcategory participation (donor/volunteer/leadership/library), and degree background. These factors clarified which combinations of attributes correlated most with pipeline entry.

Cluster analysis segmented the wider alumni base into groups ranging from "High-Intent Mature Donors" to "General Participants" enabling targeted outreach to clusters with both strong institutional ties and latent giving potential.

Predictive modelling selected the Random Search-tuned XGBoost as the optimal model for participation prediction, with latest donation time emerging as the most influential feature. This indicates that recent giving history is the strongest signal of readiness to enter the legacy pipeline.

Overall, the integration of factor analysis, clustering, and predictive modelling offers complementary perspectives: latent structure discovery, actionable segment profiling, and high-accuracy prediction. Together, these insights form a robust, data-driven foundation for targeted legacy marketing and alumni engagement strategies.

1.3 Outline

This report is developed in collaboration with LSE PAGE, in response to their growing strategic focus on legacy giving. The aim is to provide actionable, data-driven insights to support PAGE in understanding and identifying the most promising legacy prospects within LSE alumni. The report is structured to address PAGE's research questions and milestones in a staged and methodologically rigorous manner.

Literature Review is initialised to summarise relevant analytical techniques used in philanthropic research, including predictive modeling (e.g., logistic regression, decision trees), factor analysis for uncovering latent motivations, and cluster analysis for alumni segmentation. These methods are evaluated in terms of their suitability for addressing the core questions posed by PAGE.

The methodology section is structured around three core analytical approaches: factor analysis, and cluster analysis, and predictive modeling.

Data Description and EDA includes data cleaning, sanity check, and feature engineering. Then exploratory statistics of the overall alumni base and the legacy pipeline, visualisations of variable distributions is presented.

Modelling contains the intensive study of two topis: one for potential donors- alumni already in legacy pipeline that may make legacy donation, another for potential particcipants- alumni yet in legacy pipeline that may engage later. These two topics are modelled in three different approaches.

Firstly, factor analysis techniques are performed to identify latent attitudinal or behavioral dimensions that influence alumni engagement with legacy giving. Both exploratory and confirmatory approaches are used to uncover underlying structures and test theoretical

assumptions.

Secondly, clustering techniques are implemented to segment the alumni population based on observed characteristics and inferred engagement patterns. These segments will help to support the development of targeted communication and outreach strategies.

Thirdly, supervised learning models - traditioinal models (Logistic Regression with and without L1 penalty), tree-based models (Decision Tree, Random Forest, Gradient Boosting Decision Tree, and XGBoost), and Bagging methods (with Decision Trees and Logistic Regression respectively) are applied to predict the likelihood of alumni either to become a potential donor or a potential participant in the legacy pipeline. Thorough evaluations across metrics in different dimensions are conducted. The result also provides insight into the relative importance of key features.

Together, these methods provide a comprehensive framework for understanding the alumni legacy pipeline from multiple analytical perspectives, and forming the basis for data-driven recommendations later in this report.

Finally, a conclusion across all models is summarised to support more targeted legacy marketing strategies and inform future data collection efforts. A technical appendix and code handover plan are also included in accordance with Milestones 3 and 4.

2 Literature Review

2.1 Factor Analysis

Factor analysis is one of the most widely used research methods in the marketing and promotion of a programme or a product, including propagating the legacy donation. Stewart (1981) clarifies the correct applications of factor analysis in marketing including reducing dimensionality of variables, exploring data structure, as well as, testing hypotheses on any factor structures. The author also emphasizes that factor analysis should not be confused with clustering analysis and used for segmentation of potential participants, and correlation between variables should not be mixed with association when interpreting the results. In Green and Webb (1997), an exploratory factor analysis (EFA) method to discover hidden patterns that influence individuals' decisions to make monetary donations to charitable organizations, in which principal component analysis (PCA) is used for factor extraction, and the results are rotated using Varimax rotation to improve interpretability. With established criteria on eigenvalue and factor loadings, eventually six important factors are identified based on 35 attitudinal observed variables obtained from surveys. This framework is standard and widely-used in different applications. Kolhede and Gomez-Arias (2022) adopts a similar methodology with EFA, PCA, and Varimax rotation to discover factors that affect individual donation behaviours to charitable organisations. Using the Kaiser's criterion, seven factors are retained, which contributes to 58.2% of the total variance.

It is notable that PCA can be used for factor extraction in factor analysis but it is sometimes misused as factor analysis. Both Cudeck (2000) and Yang (2005) emphasize the difference between PCA and common factor analysis that PCA aims to reduce dimensionality by transforming observed variables into a smaller set of components that capture the total variance, whereas in factor analysis one should focus on common variance shared among variables, which composes the total variance with unique variance and error variance. Yang (2005) suggests that PCA should be the default option for factor extraction and researchers need to compare different available methods, including principal axis factoring, least square, maximum likelihood, alpha factoring, and image factoring, based on the context of applications. This article also summarises the results that oblique rotations generally leads to simple and interpretable results and are more realistic than orthogonal rotations since the former allows correlation between variables, although orthogonal rotations, especially the Varimax approach, are more widely-used.

In contrast to EFA, which identifies key factors, another type of factor analysis is confirmatory factor analysis (CFA), which tests whether the observed items are associated with specific factors. Sarmento and Costa (2019) applies the CFA method to test the hypothesis that the data is under a four-factor model specified by the results of EFA, in which parameters are estimated using maximum likelihood (ML) method, and various tests, such as Chi-square, comparative fit index, and Tucker-Lewis index, demonstrate a good fit of the model. This frequentist method assumes all variables are continuous even if they are ordinal or binary, while using a Bayesian framework could loosen this assumption. Ansari and Jedidi (2000) proposes a a Bayesian approach to handle multilevel binary data in which conventional max-

imum likelihood methods are computationally intractable. In this study, a multilevel normal factor model is assumed for the latent continuous variable for hierarchical data, and Markov Chain Monte Carlo (MCMC) is used to estimate model parameters, resulting in a robust, accurately recovered two-factor model. Quinn (2004) also develops a Bayesian factor analysis model that can simultaneously cope with ordinal and continuous response variables, which is common in real-world datasets, using MCMC with Gibbs sampling and Metropolis-Hastings algorithms for parameter estimation. A comparison to classical factor analysis method is conducted through a case study on political-economic risk in 62 countries in this research, which demonstrates similar rankings produced by both models but more interpretable results generated using the Bayesian method. In addition to challenges in handling complex data types, the ability to detect model misspecifications of CFA models is also of great concern. In Önen (2019), a comparison of frequentist and Bayesian approaches on identifying incorrectly omitted cross-loadings is illustrated by testing three models, CFA using maximum likelihood (ML-CFA), Bayesian CFA with noninformative priors, and Bayesian Structural Equation Modeling (BSEM), on simulated continuous data. The study concludes that the ML-CFA is more sensitive to minor misspecfications, in which the magnitude of omitted cross loading is small and less influencing, but blind for major misspecifications, while two Bayesian models are more effective for major misspecifications detection than for minor ones, and BSEM is especially suitable for complex models with high cross-loadings.

2.2 Cluster Analysis

Advances in data analysis have made a profound difference to alumni engagement and fundraising activity at universities and colleges. Cluster analysis as the top unsupervised machine learning algorithm has emerged as a significant method of segmenting alumni groups according to behavior and demographic characteristics and thereby initiating more focused and effective engagement strategies.

Le Blanc and Rucks (2009)Le Blanc and Rucks (2009) conducted cluster analysis on 33,000 records of alumni and discovered six distinct clusters with varying attributes. According to their report, a segment of alumni gave disproportionately high numbers of significant gifts, and use of clustering methods was being hailed as the answer to streamlining fundraising efforts. Durango-Cohen et al. (2012)L. and W. (2012) applied the clusterwise linear regression model to analyze alumni giving behavior and concluded that different segments of alumni responded with varying sensitivity to solicitation efforts, and as a result emphasized segmented engagement strategies.

In the area of alumni relations management, Rattanamethawong et al. (2016)Rattanamethawong et al. (2016) developed a new model based on the use of cluster methods to better understand the behavior, way of life, and features of alumni. They conducted their research with respect to the role of personalized communications and engagement strategies in strengthening alumni-institution relationships. Pedro et al. (2020)Pedro et al. (2020) segmented the alumni based on commitment and investigated the determinants of the intention to collaborate with their university and found that institutional commitment and pleasant experiences during the time spent at the university contributed significantly to alumni engagement.

Clustering application is more than traditional alumni involvement, as shown through the sophisticated use of clustering methods by Lin and Chang (2021)Lin and Chang (2009), for segmentation among online donors in Taiwan. They demonstrated that the use of RFM (Recency, Frequency, Monetary) measure and socio-demographic data increased the effectiveness in segmentation among donors for the generation of target-based promotional programs for enhancing donor retention and rate of giving.

Together, the studies demonstrate the power of cluster analysis to segment alumni groups in a manner enabling institutions to target engagement and philanthropy efforts with greater precision. With data-based approaches like these, PAGE will be better positioned to distinguish alumni attitudes and behaviors, establish stronger relationships and increased philanthropy.

2.3 Predictive Models

Predictive modeling aims to estimate the likelihood of a target outcome based on observed predictor variables. In the context of educational fundraising, these models can identify patterns that distinguish donors from non-donors, enabling targeted engagement strategies. This section reviews the main predictive modeling approaches employed in our study, grouped into four categories: traditional models, tree-based models, gradient boosting methods, and bagging methods.

2.3.1 Traditional Models

Logistic Regression Logistic regression is a widely used classification method, particularly suited for binary outcomes. Unlike linear regression, which models a continuous dependent variable, logistic regression models the log-odds of the probability that an observation belongs to a given class:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

where p is the probability of the positive class. James et al. (2013) highlight its interpretability and robustness in applied contexts such as social sciences and medicine. Hastie et al. (2009) provide a statistical treatment of logistic regression as a special case of the generalized linear model, typically estimated via maximum likelihood.

Logistic Regression with L1 Regularization L1-regularized logistic regression, also known as Lasso logistic regression, introduces a penalty term $\lambda \sum_{j=1}^{k} |\beta_j|$ to the likelihood function. This encourages sparsity in the coefficients, performing variable selection in the presence of correlated predictors (Tibshirani, 1996). This is particularly valuable when dealing with high-dimensional alumni engagement data, where many predictors may be redundant.

2.3.2 Tree-Based Models

Decision Trees Decision trees are non-parametric models that recursively partition the predictor space into disjoint regions. The Classification and Regression Tree (CART) algo-

rithm (Breiman et al., 1984) selects splits that maximize reduction in impurity, measured by criteria such as the Gini index or cross-entropy. While decision trees require minimal preprocessing and offer clear interpretability, they are prone to overfitting when fully grown. Pruning methods (Hastie et al., 2009) mitigate this by balancing fit with model complexity.

Random Forests Random forests (Breiman, 2001) extend decision trees through bootstrap aggregation (bagging) and random feature selection. Each tree is trained on a bootstrapped sample of the data, and at each split, a random subset of features is considered. This decorrelation among trees reduces variance and improves generalization. Although interpretability is reduced compared to a single decision tree, random forests are highly effective for structured tabular data.

2.3.3 Gradient Boosting Methods

Gradient Boosted Decision Trees (GBDT) GBDT (Friedman, 2001) builds an ensemble of trees sequentially, with each new tree trained to fit the residual errors of the previous ensemble using gradient descent on a differentiable loss function. This method often achieves higher accuracy than bagging approaches but requires careful regularization to avoid overfitting.

XGBoost XGBoost (Chen and Guestrin, 2016) is a scalable, regularized implementation of gradient boosting that incorporates second-order gradient information, L1/L2 regularization, and advanced tree-pruning techniques. It is known for its efficiency on large datasets and has been widely adopted in machine learning competitions.

2.3.4 Bagging Methods

Bagging with Decision Trees Bootstrap aggregation, or bagging (Breiman, 1996), reduces model variance by averaging predictions from multiple base learners trained on bootstrapped datasets. When decision trees are used as base learners, bagging can significantly improve predictive stability while retaining the trees' flexibility.

Bagging with Logistic Regression Bagging can also be applied to more stable base learners like logistic regression, although the variance reduction benefits are less pronounced. However, in high-variance settings (e.g., when regularization is weak or the dataset is small), bagging logistic regression can still yield performance improvements.

3 Methodology

3.1 Factor Analysis

Factor analysis an unsupervised, non-parametric statistical method which explains variability among observed variables in terms of some latent variables, called factors. The main idea is to seek for the mutual dependence of these variables on, usually a fewer number of, unobserved factors. Suppose there are P observed variables each with n data points $X_{p,i}$, where $p = 1, 2, ..., P$ and $i = 1, 2, ..., n$, and K common factors $F_{k,i}$,where $k = 1, 2, ..., K$ and $K < P$. Then, each variable is expressed as a weighted linear combination of these factors and an error term, which is provided below.

$$
\begin{aligned}
& X_{p,i} - \mu_p = \lambda_{p,1} F_{1,i} + \lambda_{p,2} F_{2,i} + \cdots + \lambda_{p,k} F_{K,i} + \epsilon_{p,i} \\
\text{where } & \lambda p, k \text{ is the loading for the variable } X_p \text{ on factor } F_k, \\
& \mu_p \text{ is the mean of } X_p, \text{and} \\
& \epsilon_{p,i} \text{ has zero mean and finite variance.}
\end{aligned}
$$

This suggests that variables are not independent and driven by some common factors. Loadings are coefficients that describe the relationship between observed variables and latent factors, in which a high loading indicates a strong association. The model also assumes that factors and errors are independent, i.e. $Cov(F_k, \epsilon_p) = 0$, $E[F_k] = 0$ and $Var(F_k) = 1$, so $E[X_p]$ is subtracted from $X_{p,i}$ in the equation to satisfy the zero mean assumption. By defining $\mathbf{\Sigma} = Cov(\mathbf{X}) \in \mathbb{R}^{P \times P}$, the total variance among variables can be decomposed to the sum of common variance shared across variables and unique variance for each variable. This allows the representation of higher dimensional variables by lower dimensional latent factors and has the following mathematical formulation.

$$
\begin{aligned}
& \mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top + \mathbf{\Psi} \\
\text{where } & \mathbf{L}\mathbf{L}^\top \text{ quantifies the shared variances, and} \\
& \mathbf{\Psi} \in \mathbb{R}^{P \times P} \text{ is the diagonal matrix of unique variances}
\end{aligned}
$$

Factor analysis is commonly used for investigating the underlying dimensions for events or phenomenons and validating and testing the constructed factor structure and relationship between variables. The two main techniques of factor analysis can achieve such purposes: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The former aims to identify the hidden factors given observed data through estimating factor loadings, and generate hypotheses related to the latent structure, whereas, the latter examines the proposed, either by EFA or not, model given both data and latent factors.

3.1.1 Exploratory Factor Analysis

Extraction and Rotation Coefficients of hidden factors are first estimated via extraction methods such as principal components analysis (PCA) and principal axis factoring (PAF). PCA is designed to maximise explained variance in the observed variables, and it accounts

for 100% of variance (including unique and error variance). Hence, PCA is more dimension-reduction-focusing, though it has been widely used in EFA. While, PAF is a common factor analysis method that extracts factors based on shared variance, which is flexible since it does not assume certain distribution of data. The number of factors should be used can be selected by different methods. Eigenvalues represent the amount of variance explained by each factor, and the Kaiser Criterion retains factors with eigenvalues greater than or equal to 1, as these factors could explain more variance than others. The scree plot can obtained by plotting eigenvalues against factor numbers, and the elbow rule suggests that factors above the point where the curve bends sharply should be retained in the model. To interpret estimated loadings, the loading matrix $\mathbf{L}$ is rotated to demonstrate the relationship between each factor with groups of variables. There are two commonly used types of rotations, orthogonal and oblique. Orthogonal rotation remains factors to be uncorrelated and is simple to implement and interpret, such as Varimax rotation. While oblique rotation, such as Promax rotation, might be more suitable when factors are correlated, and so is more realistic in real-world applications.

Evaluation Before performing EFA, the suitability of the data for factor analysis can be evaluated using the following measures. Kaiser-Meyer-Olkin (KMO) Measure assesses sampling adequacy, with values above 0.6 considered acceptable for EFA. Moreover, Bartlett's Test of Sphericity determines whether the correlation matrix is suitable for factor analysis, where a significant (p-value smaller than 0.05) result indicates sufficient correlations among variables. Variance Inflation Factor (VIF) is used to detect multicollinearity, with VIF greater 10 suggesting severe issue in data. After models being fitted, the results can be evaluated in different aspects. Features with factor loadings greater than 0.4 can be retained to ensure meaningful associations with their respective factors. Communalities represent the proportion of variance in each variable explained by the extracted factors, with values above 0.5 indicating adequate representation. The reliability of factor scores can be assessed by factor determinacy coefficients, in which values greater than 0.8 are considered highly reliable. The total variance accounted for by the retained factors is called cumulative variance explained, whose thresholds varying by discipline but typically exceeding 60%.

3.1.2 Confirmatory Factor Analysis

In CFA, the proposed latent structure indicates zero loadings for specific factors and allows other non-zero loadings to be estimated. The general model-implied covariance matrix, which relaxes the assumptions of uncorrelated, standardised factors stated above, is defined as

$$\begin{aligned} \boldsymbol{\Sigma}(\theta) &= \mathbf{L}\boldsymbol{\Phi}\mathbf{L}^\top + \boldsymbol{\Theta}_{\boldsymbol{\epsilon}} \\ \text{where } \boldsymbol{\Phi} &= Cov(\mathbf{F}) \text{ is the covariance matrix of factors, and} \\ \boldsymbol{\Theta}_{\boldsymbol{\epsilon}} &= Cov(\boldsymbol{\epsilon}) \text{ is the diagonal matrix of unique error variances} \end{aligned}$$

Hence, the parameters $\theta = \{\mathbf{L}, \boldsymbol{\Phi}, \boldsymbol{\Theta}_{\boldsymbol{\epsilon}}\}$ are selected to as close as possible to the sample covariance matrix, $\mathbf{S}$, based on the data, and can be estimated through different methods,

such as maximum likelihood, Laplace approximation, and MCMC. The quality of model fit can be assessed by various measures. Chi-square to Degrees of Freedom Ratio measures the discrepancy between the sample covariance matrix and model-implied covariance matrix, adjusted for model complexity, and values between 1-3 indicate a good fit. Comparative Fit Index (CFI) suggests the relative improvement in fit of the hypothesised model compared to a null (independence) model. The error of approximation per degree of freedom, accounting for model parsimony is measure by the Root Mean Square Error of Approximation (RMSEA), with a value below 0.05 indicating a good fit.

3.2 Cluster Analysis

This section outlines a plan to conduct an unsupervised learning analysis of LSE alumni data to identify underlying clusters of behaviour and engagement. Clustering serves as a data-driven alternative to manual segmentation, offering the ability to detect hidden structure and latent engagement patterns in high-dimensional, complex data. It provides a foundation for describing behavioural archetypes, understanding the factors associated with progression through the legacy pipeline, and ultimately supporting more targeted and efficient fundraising strategies.

Given the scale of the data and the diversity of features, this study proposes a comparative evaluation of three clustering methods: k-means, hierarchical agglomerative clustering, Bayesian Gaussian mixture models (BGMM) and DBSCAN. Each of these methods offers distinct advantages, and their suitability will be evaluated based on internal validation metrics, cluster interpretability and practical usability.

K-means clustering is a natural first step due to its scalability and interpretability. It partitions data into k mutually exclusive clusters by minimising within-cluster variance. K-means performs well when clusters are spherical and evenly sized—assumptions that may hold for many behavioural traits such as frequency of alumni engagement or donation participation. K-means' speed makes it suitable for running on the full dataset, enabling us to capture broad engagement patterns.

However, alumni engagement patterns may not conform to simple spherical clusters. Therefore, hierarchical clustering will also be implemented, which builds a nested tree of clusters by iteratively merging the most similar pairs based on a linkage criterion. This approach does not require pre-specifying the number of clusters, allowing us to explore different levels of granularity by cutting the dendrogram at various heights. Although it is computationally more expensive, it provides valuable visual insights through dendrograms and is useful for exploratory analysis.

A third technique proposed is Bayesian Gaussian mixture modelling. The Bayesian extension introduces a Dirichlet process prior that enables the model to infer the optimal number of components automatically, avoiding the need for hard-coded k values. Also, BGMM is potentially usable as inputs to downstream probabilistic models such as discrete-time hazard models or Markov transition systems. A little more computationally intensive, BGMM is a good method for handling overlapping, non-spherical behavioural trajectories and uncertainty in cluster allocation.

The last technique is DBSCANDensity-Based Spatial Clustering of Applications with Noise. DBSCAN does not rely on presetting the number of clusters, but can discover the cluster structure in any shape and identify the noise. Concerning complex alumni data in legacy donation, DBSCAN is suitable for capturing the edge group and local high density area. DBSCAN does not hypothesize the distribution of data, so it is well-adapted in dealing with outliers and not normal distribution. This is a good supplement method to explore our understanding of the legacy alumni data.

To compare and evaluate clustering solutions, internal measures like cluster compactness will be calculated for raw samples to verify robustness and stability of the cluster solution. Following the clustering process, every cluster will be described with average feature values, demographic profile, and behavior, determining significant segments like "General Participants" or "High-Potential Alumni".

At the integration point, the memberships in clusters can be embedded in the tools and dashboards of Advancement, for example, to target mail communications or priorities for donor stewardship. High-potential clusters, for example, may be highlighted as being deserving of proactive solicitation, and low-return segments deprioritized. The product of this analysis will drive targeting strategy, donor cultivation, and pipeline projections, and guide toward a data-based and more effective advancement operation.

3.3 Predictive Models

3.3.1 Traditional Models

Logistic Regression We implement logistic regression for binary classification, modeling the log-odds of donation probability as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j,$$

where $p = P(y = 1 \mid \mathbf{x})$ denotes the probability that an alumnus donates, and x_j are the predictor variables. Parameters are estimated via maximum likelihood estimation (MLE). Continuous predictors are standardized to mean zero and unit variance, while categorical variables are one-hot encoded. No regularization is applied in this baseline model, providing a benchmark for interpretability and coefficient analysis.

Logistic Regression with L1 Regularization The L1-regularized logistic regression, or Lasso logistic regression, adds a sparsity-inducing penalty term to the likelihood:

$$\hat{\beta} = \arg\min_{\beta} \left\{ -\ell(\beta) + \lambda \sum_{j=1}^{k} |\beta_j| \right\},$$

where $\ell(\beta)$ is the log-likelihood function and λ controls the degree of regularization. This encourages variable selection, which is particularly beneficial for high-dimensional alumni engagement data. The regularization parameter λ is selected via cross-validation.

3.3.2 Tree-Based Models

Decision Trees We train Classification and Regression Trees (CART) using the Gini index as the impurity measure. Trees are grown until a minimum split size is reached, followed by post-pruning to avoid overfitting. Features require minimal preprocessing, and categorical variables are label-encoded. Decision trees offer intuitive decision rules that can be easily interpreted by non-technical stakeholders.

Random Forests Random forests combine multiple decision trees trained on bootstrapped datasets, with a random subset of features considered at each split. This decorrelation between trees reduces variance and improves generalization. In our context, random forests also provide feature importance measures, offering insights into which engagement metrics most strongly influence donation likelihood.

3.3.3 Gradient Boosting Methods

Gradient Boosted Decision Trees (GBDT) GBDT models are trained sequentially, with each tree fitting the residuals of the previous ensemble. We use a logistic loss function for binary classification, optimizing via gradient descent. Hyperparameters such as learning rate, maximum tree depth, and number of estimators are tuned using validation data to balance bias and variance.

XGBoost XGBoost extends GBDT with second-order gradient information, L1/L2 regularization, and optimized tree-pruning strategies. Missing values are handled natively during split finding. Its scalability and regularization make it particularly effective for large alumni datasets with mixed variable types.

3.3.4 Bagging Methods

Bagging with Decision Trees Bagging involves training multiple decision trees on different bootstrap samples of the training data and aggregating their predictions via majority voting. This method reduces variance, particularly for high-variance learners like decision trees, and improves predictive stability.

Bagging with Logistic Regression Although logistic regression is a relatively low-variance model, applying bagging can still offer improvements in unstable settings—such as when the dataset is small or the model is weakly regularized. We aggregate the predictions from multiple logistic regression models trained on bootstrap samples to form the final prediction.

3.3.5 Model Evaluation

All models are evaluated using stratified hold-out validation to preserve class proportions. Evaluation metrics include accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC), enabling a balanced assessment of predictive performance across models.

4 Data Description and EDA

In this section, an end-to-end workflow for the data exploration and preprocessing is outlined prior to modeling, following a structured sequence: **data merging → sanity check → feature engineering → label analysis**. Specifically, for Potential Donors, **legacy donation pipeline analysis** is subsequently performed.

There are seven original datasets:

- ANON_LegacyEnquirer_20250203.csv, abbreviated with *Enquirer*,
- ANON_Legacy_Intender_20250130.csv, abbreviated with *Intender*,
- ANON_LegacyPledge_2050130.csv, abbreviated with *Pledge*,
- ANON_LegacyProposals_20250206.csv, abbreviated with *Proposal*,
- ActiveAlum_Anonymised_20250205.csv, abbreviated with *ActiveAlum*,
- ANON_LegacyCircle_20250130.csv, abbreviated with *Circle*, as well as,
- Individual Data Points.xlsx, refered as *Individual Data Points* later in this report.

The first five datasets records IDs and information of alumni who have participated in each stage of the legacy pipeline or who have joined in the alumni circle. *ActiveAlum* contains demographic features of alumni who are actively connected to the School and the PAGE department. *Individual Data Points* is another collection of information for alumni relation to the School for past years, where some IDs coincide with the ones in *ActiveAlum*.

A overview of the first six datasets is demonstrated as below. *Pledge* has the largest number of columns and the largest proportion of invalid values, including meaningless zero values, empty strings, symbols and missing fields. *Individual Data Points* contains eight sheets corresponding to years from 2017 to 2024, where columns in each sheet are the same. The percentage of invalid values in each sheet is between 60% and 64%. To address the two main research questions mentioned in section 1.1, two tables are created from the above seven datasets given as described in following subsections.

Table 1: Original Datasets Overview

Datasets	**Rows**	**Cols**	**Invalid Values**	**Numeric Cols**	**Categorical Cols**
Enquirer	324	11	38.83%	3	8
Intender	123	9	26.47%	1	8
Pledge	191	29	67.25%	2	27
Proposal	738	10	14.20%	1	9
Circle	234	9	32.95%	1	8
ActiveAlum	192168	10	38.92%	3	7

4.1 Dataset 1: merged_pipeline

To explore the characteristics of existing legacy donors against other alumni who have joined the legacy pipeline, the first dataset **merged_pipeline** is created, containing all IDs in the pipeline.

4.1.1 Data Merging

The four engagement-related datasets are integrated first: *Enquirer*, *Intender*, *Pledge*, and *Proposal*. Tables are merged by outer joins with 4 separate columns indicating the existence of each unique ID in each table. The integration is performed as follows:

- The unified primary key is standardized as `ID_NUMBER` across all datasets, and data types are harmonized.
- For alumni IDs appearing multiple times in either *Enquirer* or *Proposal*, only the most recent record (based on `COMPLETED_DATE` and `START_DATE`) is retained, along with all relevant attributes such as `COUNTRY_CODE`.
- After deduplication by `ID_NUMBER`, boolean flags are created to indicate whether an individual appears in each of the four datasets (`stage_enquirer`, `stage_intender`, `stage_pledge`, `stage_proposals`).
- For similar columns, such as `CITY` originally in *Intender* and `P_CITY` in *Pledge*, absence of conflicts in values is confirmed, and only one of columns of the same type is retained after combining them.
- Any columns with more than 90% invalid values are removed.

Next, the alumni profiling datasets *ActiveAlum* and *Circle* are merged to enrich user attributes. Among 875 IDs in merged version of above four tables, 647 are not in *Circle* and 430 are not in *ActiveAlum*. For the *Individual Data Points* dataset, two new features for each ID are created based on columns in the table:

- **sheet_appear_count**: the number of unique years IDs appear in the entire dataset between 2017–2024.
- **subcategory_Y_count**: the number of unique years IDs were active in any given `sub_category` columns during the same period.

Any IDs with non-missing values in the column `PROPOSAL_TYPE_ECHO` are considered to be donors of legacy and others are marked as non-donors, which is recorded in the column `label` with value 1 for donors and value 0 for non-donors.

4.1.2 Feature Engineering

Before modeling, a series of sanity checks are conducted to ensure data validity and to remove variables that are either irrelevant or too sparse for predictive modeling(more on submitted notebook, see 1.4 for sanity check and feature engineering). After the sanity check, feature engineering is performed to ensure consistent and meaningful representations for modeling:

- **Date variables:** For all date or datetime fields (excluding `degree_year`), the year difference relative to the current year of 2025 is calculated. Missing values are then replaced with 0.

- **Age estimation and grouping:** Missing `AGE` values are estimated using graduation year information (`graduation_annum` = 2025 - `DEGREE_YEAR`) when available. By default, the estimated age is calculated as 21 + `graduation_annum`. However, for individuals with `Bequest Notification` records, the age estimation is calculated as: 21 + `graduation_annum` - 2025 + `STOP_DATE` year. When both `AGE` and `graduation_annum` are missing, values remain as missing. Age is then binned into quantile-based groups, with missing values labeled as `"Unknown"`.

- **Numeric variables:** For other `int64` fields, missing values are also encoded as 0 after data validation.

- **Categorical variables:** Missing or empty strings in object-type fields are replaced with `"Unknown"`. Categories with fewer than 50 occurrences are merged into a single `"Other"` category to avoid extreme values. This ensures that each categorical variable has a manageable number of 2–5 categories, which is suitable for models like logistic regression or decision trees under a sample size of `875`, thus reducing potential overfitting.

Following the sanity check and feature engineering, a refined dataset that will be used for modeling and later analysis is obtained. Due to the existence of many missing values, all unique ID records are retained and missing values have been processed.

4.1.3 Label Analysis

Up to this point, `merged_pipeline` contains all features that are considered to be helpful to distinguish donors of legacy from the alumni population. To demonstrate the distributions of features among donors and non-donors in different aspects, two temporary variables are created to classify alumni according to if a donor has passed away and left the legacy to the School and if an alumnus has joined the pipeline:

- `legacy_BN`: takes value `BN` if `PROPOSAL_TYPE_ECHO` is `Bequest Notification`, indicating the School has been left with the legacy; takes value `Not_BN` if `PROPOSAL_TYPE_ECHO` is any other non-Missing status, suggesting a firm commitment is made to leave a legacy; and takes value `Missing` if `PROPOSAL_TYPE_ECHO` is `Missing` and proposals to donate have not been made.

- `pipeline_donate`: takes value `Both` if the person has joined the legacy pipeline (at least one stages among *Enquirer*, *Intender*, and *Pledge*) and became a donor (either left legacy to the School or committed to do); takes value `No_donate` if the person has joined the pipeline but has not donated (including committed to donate); and takes value `No_pipeline` if the person has made donation without participating in any of the three stages.

Among 875 alumni in `merged_pipeline`, 66.9% of them have completed donation to the School or committed to leave a legacy, which are called donors, and the rest 290 people have joined in at least one of the stages of *Enquirer*, *Intender*, and *Pledge* but have not decide to donate. Of 585 donors, 40.3% of them (236) have passed away and left a legacy, and 67.7% (396) have not joined the legacy pipeline. Among donors who have not participated in the pipeline, more than half are marked as Bequest Notification, and 23.7% of them have made verbal or written proposals. 60.5% of alumni in the legacy pipeline have missing `PROPOSAL_TYPE_ECHO`, i.e. have not made donation, and only 4% of pipeline participants have left the legacy to the School already.

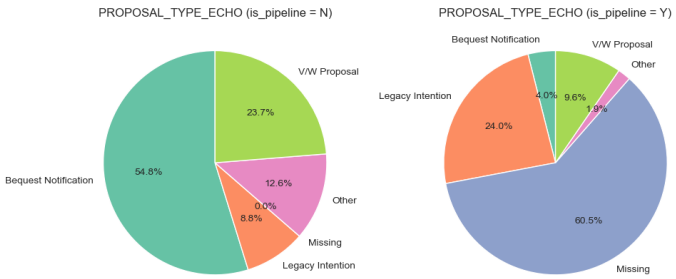


Figure 1: Pie Chart of PROPOSAL_TYPE_ECHO on is_pipeline

The distributions of features on `label`, `legacy_BN`, and `pipeline_donate` are investigated via summary statistics and plotting as below. Features are categorised into 3 types: demographic features, pipeline-related features, and alumni-relation features.

Demographic Features Age is one of the most discriminating features among all. By drawing boxplots of `age_extend` on `label`, `legacy_BN`, and `pipeline_donate`, it can be observed that donors have generally higher ages than non-donors, and among those donors, fewer senior alumni have participated in the legacy pipeline. It is not surprising that donors who are holding commitment to donate are generally of lower ages than those who have left legacy already.

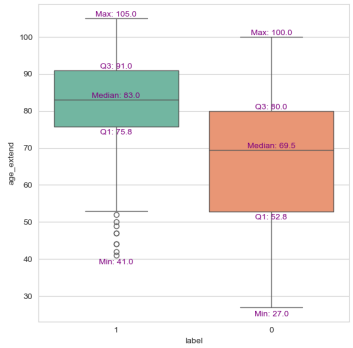


Figure 2: Boxplot on label and age_extend

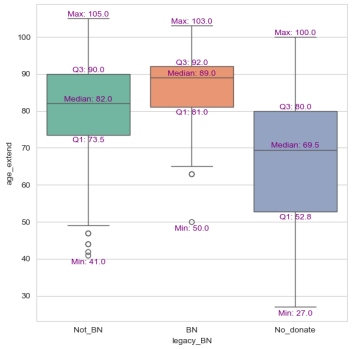


Figure 3: Boxplot on legacy_BN and age_extend

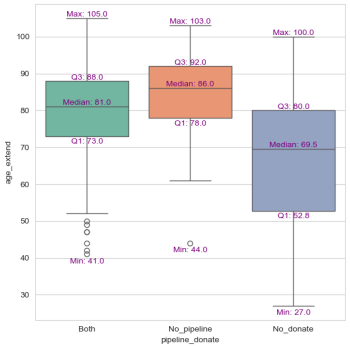


Figure 4: Boxplot on pipeline_donate and age_extend

`graduation_annum` is the number of years from graduation until the current year for

alumni, which is positively correlated to `age_extend`. Thus, the boxplots of these two features demonstrate similar distributions in different groups of people in the dataset.

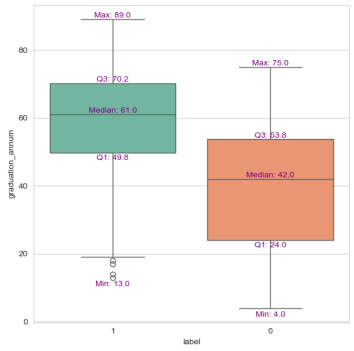


Figure 5: Boxplot on label and graduation_annum

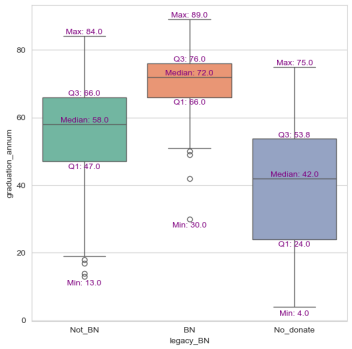


Figure 6: Boxplot on legacy_BN and graduation_annum

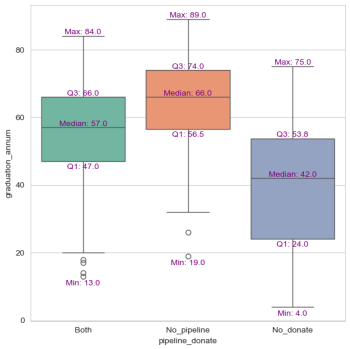


Figure 7: Boxplot on pipeline_donate and graduation_annum

Due to a large amount of missing values in `Marital Status`, no obvious pattern could be observed across different types of labels. An alternative feature reflecting family-related information is `Children Y/N`, reflecting if alumni have children or if values are missing in this field. 49.1% of values in this field is Unknown, and 42.3% of the alumni recorded in `merged_pipeline` do not have children. More non-donors do not have children than donors, probably because of they are of younger ages. The large amount of missing values in this field for donors is mainly caused by alumni who have left legacy already and have not left enough personal information to the School.

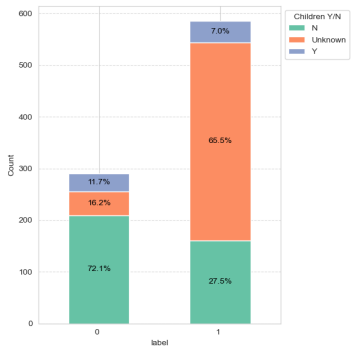


Figure 8: Barplot on label and Children Y/N

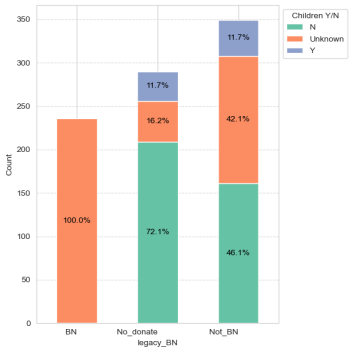


Figure 9: Barplot on legacy_BN and Children Y/N

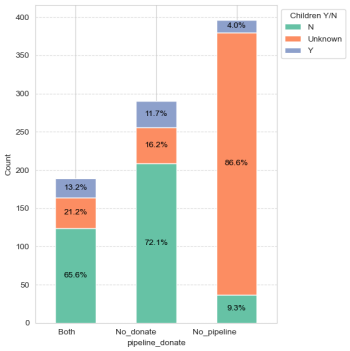


Figure 10: Barplot on pipeline_donate and Children Y/N

The distribution of missing values in `COUNTRY_CODE` is similar to that in `Children Y/N`. Ignoring unknown `COUNTRY_CODE` records, countries of origin of alumni in each group are dominated by the United Kingdom, as demonstrated below.

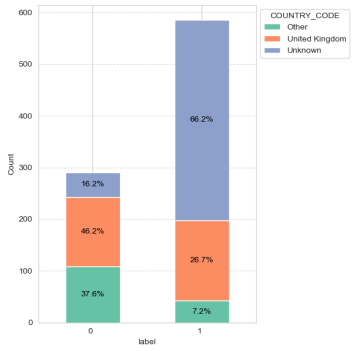


Figure 11: Barplot on label and COUNTRY_CODE

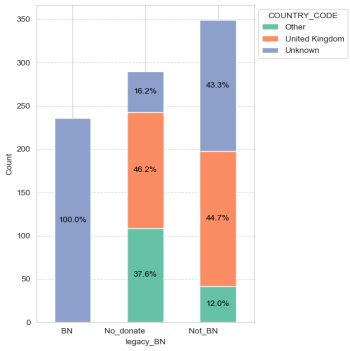


Figure 12: Barplot on legacy_BN and COUNTRY_CODE

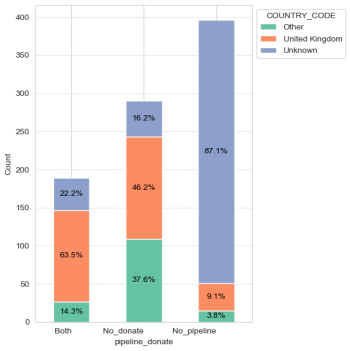


Figure 13: Barplot on pipeline_donate and COUNTRY_CODE

Pipeline-related Features `enquire_annum` and `proposal_annum` measure the numbers of years from the time of making enquiry or proposal, respectively, to the current year, where missing values are recorded as 0. The left plot below shows that most non-donors have positive `enquire_annum` with median of 3.5 years, and oo few donors who have positive values in this field to draw a box. The other two plots are drawn on `proposal_annum` with `legacy_BN` and `pipeline_donate`. Donors who have not leave the legacy have much longer `proposal_annum` than who are noted as Bequest Notification, while donors who have joined the pipeline have similar `proposal_annum` values with who have not. In addition, non-donors have no proposal records by definition.

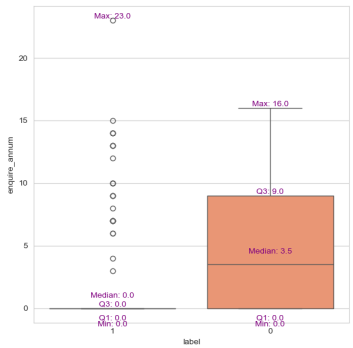


Figure 14: Boxplot on label and enquire_annum

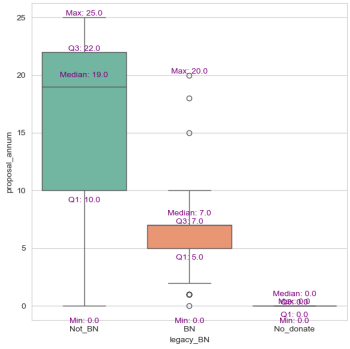


Figure 15: Boxplot on legacy_BN and proposal_annum

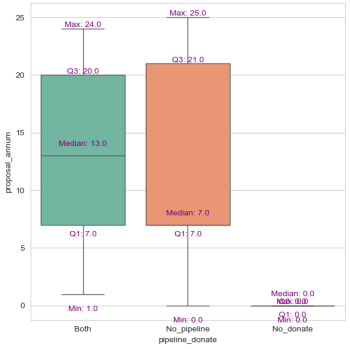


Figure 16: Boxplot on pipeline_donate and proposal_annum

There exist multiple enquiring and proposal records of the same ID, and the numbers of records are stored in `enquirer_count` and `proposals_count`. The left plot below indicates that most donors have 0 values in `enquirer_count`, either not entering the enquiry stage or missing this field, and a small portion of non-donors has `enquirer_count` greater than 1. Observing the rest two plots, patterns coincide and donors who have left the legacy and have not joined the three stages of the pipeline generally have higher `proposals_count` than others.

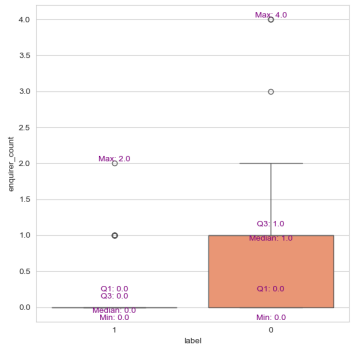


Figure 17: Boxplot on label and enquirer_count

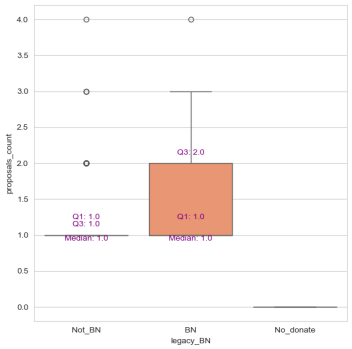


Figure 18: Boxplot on legacy_BN and proposals_count

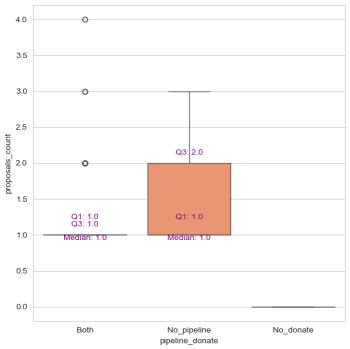


Figure 19: Boxplot on pipeline_donate and proposals_count

Alumni-relation Features `latest_donation_annum_flag` indicates if alumni have donated in any form to the School, where missing values are categorised into N. 71.4% non-donors of legacy and 28.5% donors have other types of donation to the School. Nearly half of the donors who have not passed away and two third of donors who have joined the pipeline have made donation before. The large proportion of Ns in donors is because all donors who have left the legacy and 89.4% donors who are out of the pipeline have never donated to the School in other forms. `gift_annum_flag` is a similar feature, reflecting if alumni have sent gift to the School. However, the difference between donors and non-donors in this field is not significant.

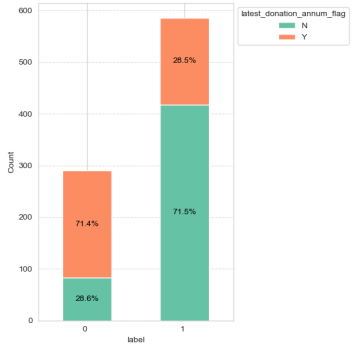


Figure 20: Barplot on label and latest_donation_annum_flag

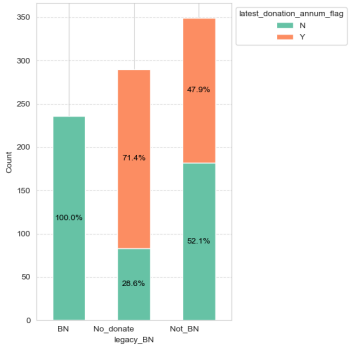


Figure 21: Barplot on legacy_BN and latest_donation_annum_flag

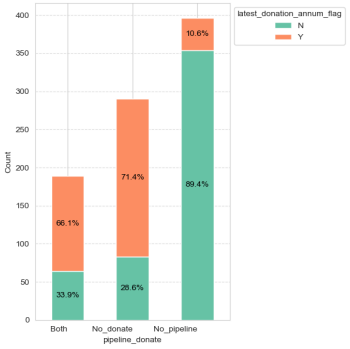


Figure 22: Barplot on pipeline_donate and latest_donation_annum_flag

There are 8 sheets in *Individual Data Point* for years from 2017 to 2024. Whether alumni appear in any sheets is recorded in `sheet_appear_count_flag`. There are 4 subcategory columns in each sheet of *Individual Data Point*: donor, volunteer, leadership group, and library. `subcategory_Y_count_flag` stores whether alumni have participated in any one of these subcategories. The distributions of these two features are highly similar, so only plots for `sheet_appear_count_flag` are demonstrated below. A higher proportion of non-donors (28.6%) than donors (13%) are present in *Individual Data Point*. All donors noted as Bequest Notification and 78.2% donors who have not passed away are absent in this dataset. More donors who are in the pipeline appear in at least one sheet than those who are out of the

pipeline.

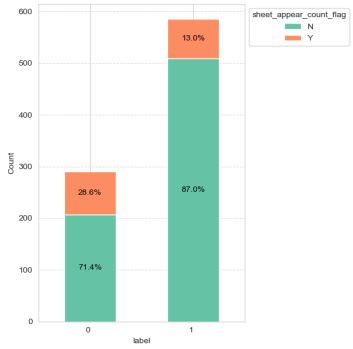


Figure 23: Barplot on label and sheet_appear_count_flag

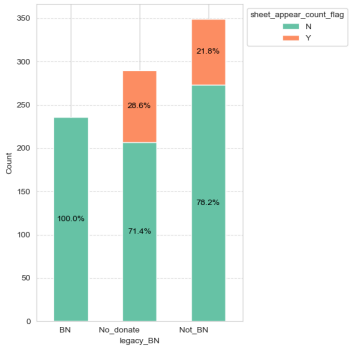


Figure 24: Barplot on legacy_BN and sheet_appear_count_flag

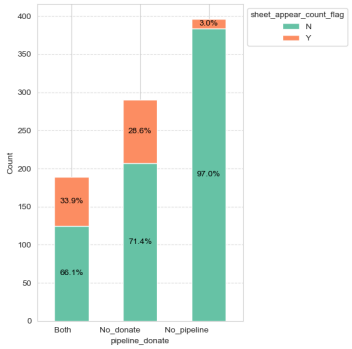


Figure 25: Barplot on pipeline_donate and sheet_appear_count_flag

4.1.4 Legacy Donation Pipeline Analysis

Preliminary Variable Distributions This is an exploratory data analysis on the distribution of alumni across different stages of the Legacy giving pathway. The data were merged from four key stages—Enquirer (initial inquiry), Intender (expressed interest), Pledge (commitment), and Proposal (formal proposal)—into a single table and counted the number of stages each alumnus participated in.

Figure 26 shows the number distribution of the alumni in the four stages of the legacy giving. The Proposal Stage has the highest number of the alumni, over 700 people, which means many alumni formally have gone through communicating and proposing procedure. Alumni in The Intender Stage are the least, only around 150 people. This shows there are part of the alumni directly converting from Enquirer to Pledge, bypassing the Intender Stage in the middle.

The legacy pipeline doesn't quite like a funnel shape, as not every alumni will go through the whole stages. It is possible to bypass a stage, or enter the pipeline at any point. Besides, the high population of alumni in Proposal should be focused on, analysing if they choose to donate finally or give up. What's more, the way to motivate the alumni in Enquirer to the further stages is also important to be discover.

Figure 27 shows the number of stages in which the alumni participate, measuring the integrity of the pipeline.

The results showed that only a small number of alumni went through all four stages, while nearly half engaged in just one stage, indicating a high drop-off rate along the pathway.

Furthermore, the stage distribution and overlap reveals that some alumni transitioned through two or three stages but did not reach the final proposal stage. This suggests that, beyond identifying new potential donors, it is crucial to focus on and motivate the "high-potential group" — those who have already shown interest but have not yet converted.

To conclude, most alumni didn't continue to participate, and the loss was serious, thus maintaining their initiatives needs attention. The middle stages (Intender, Pledge) require further follow-up.

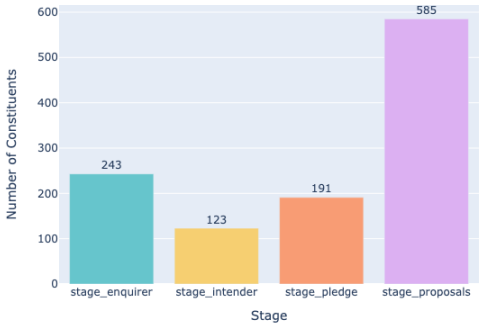


Figure 26: Number of Constituents by Stage

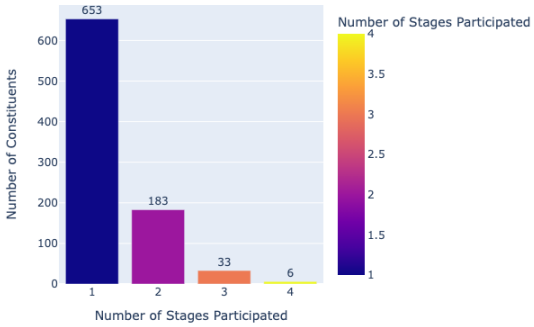


Figure 27: Distribution of Stage Participation Count

Pipeline Efficiency Insights To truly measure the effectiveness of the legacy giving pipeline, it is essential to move beyond aggregate counts and analyze the specific journeys of individuals who successfully convert into prospects. While the designed path follows a sequence of Enquirer, Intender, Pledge, and finally Proposal, the reality of donor engagement is far more fluid. Individuals can enter at any point, skip stages, or be cultivated through separate relationship management channels like the Circle. This analysis, therefore, is critical for uncovering hidden, high-performing pathways that can inform future outreach. Focusing exclusively on the population of positive samples (stage_proposal=TRUE) could dissect which engagement patterns are most fruitful and tell a data-driven story about how interest truly translates into commitment.

The first step is classifying all 875 individuals in the dataset into three behaviorally distinct groups (Full Journey, Partial Journey and No Journey) based on their engagement with the three core stages: Enquirer, Intender, and Pledge. Then the rate at which individuals in each group successfully convert by reaching the Proposal stage is calculated. This approach allows us to directly measure the return on investment for different engagement strategies.

Figure 28: Conversion Effectiveness of Each Engagement Journey

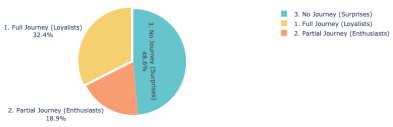


Figure 29: Distribution of Conversion Rates by Journey Type

As Figure 28 shows, the three groups are divided into two parts, respectively, on whether the individual reaches the Proposal Stage. The Partial Journey Group contains the largest number of individuals, but the conversion rate is low. While the Full Journey Group contains the smallest number of alumni, the conversion rate is much higher. This indicates the effect of the full pipeline, which means guiding alumni to follow up the whole pipeline procedure is a good choice to result in a final donation.

Figure 29 focus on the relative percentage contribution of the three groups in the overall conversion rate. 'No Journey (Surprises)" with 100% conversion rate captures the largest proportion(48.6%), which shows this group converts the most successfully without joining the pipeline. "Full Journey(Loyalists)" with 66.7% conversion rate is 32.4%. Although the alumni number of this group is small, the success of conversion is deserved to be noticed. Most importantly, Rather than relying solely on traditional pipeline optimization, future strategies should focus on identifying and expanding efficient sources of No Journey Group.

Table 2: Conversion Rate Comparison by Stage Participation

Stage Group	**Total Individuals**	**Converted to Proposal**	**Conversion Rate (%)**
Enquirer	243	33	13.58
Not Enquirer	236	156	66.10
Intender	123	64	52.03
Not Intender	356	125	35.11
Pledge	191	134	70.16
Not Pledge	288	55	19.10

Figure 30: Conversion from Two-Stage Journeys to Proposal

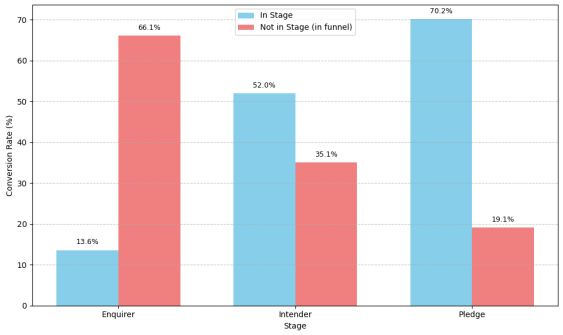


Figure 31: Proposal Conversion Rate: In Stage vs Not in Stage

Based on the results of the conversion analysis of the three core stages in the donation pipeline, total number of individuals in the Enquirer Stage was 243, but only 33 of them reach the final Proposal Stage, with merely 13.58% conversion rate, whereas those who skipped it had a much higher conversion rate of 66.10%.This suggests that early-stage enquirers may not be strong indicators of final commitment, and that some high-value prospects bypass the formal enquiry step entirely. In contrast, the conversion rate in the Intender Stage is 52.03%,, compared to 35.11% for those who skipped it. The probability of reaching final proposals is higher after an individual expresses a clear intention to donate, which is a useful indicators. The most prominent stage is Pledge with a very ideal conversion rate 70.16%, versus only 19.10% among non-participants, indicating the Pledge Stage shows strong initiatives of donating as well as great conversion tendency. Meanwhile, this stage covers a relatively large number of individuals. Therefore, in future donor cultivation strategies, emphasis should be placed on identifying, cultivating, and communicating with individuals who are at or about to enter the Pledge stage, in order to achieve more efficient resource allocation and donor transformation.

Table 3: Conversion and Drop-off Rates by Two-Stage Journey Combination

Partial Path	Total Count	Proposal Count	Conversion Rate (%)	Drop-off Rate (%)
Enquirer + Intender	22	3	13.64	86.36
Enquirer + Pledge	24	13	54.17	45.83
Intender + Pledge	14	14	100.00	0.00

Table 3 and Figure 31 show the difference among the three double-stage journeys. 'Intender + Pledge' is the most surprising combination with 100% conversion rate. This shows that the combination of Intender and Pledge is a very persuasive conversion mode, which deserves to be repeated. 'Enquirer + Pledge' has a conversion rate of 54.17%, indicating the need for further communication. "Enquirer + Intender" only ensures the initial contact and intention with firmly pledging, which has 86.36% drop-off rate. This shows the importance of the Pledge Stage in further detail. To sum up, alumni should be strongly encouraged to follow the "Intender + Pledge" path.

In conclusion, this conversion-focused analysis provides clear, data-driven guidance for resource allocation. The "Full Journey" is the gold standard of efficiency and should be considered a practice model for cultivating top-tier prospects. The "Partial Journey," especially paths centered on the Pledge stage, is the primary engine for driving conversions on a scale. Finally, the "No Journey" group represents a significant opportunity but reminds us that broad outreach must be paired with more sophisticated identification methods to effectively mine this pool for high-potential individuals. The findings also highlight the critical role of the Pledge phase. In contrast, participation in the Enquirer stage appears to be a weaker predictor of conversion and may require improved screening or prioritization to identify truly high-potential prospects. Therefore, future strategies should focus more heavily on identifying and supporting individuals who have already reached, or are on the cusp of entering, the Pledge stage.

4.2 Dataset 2: active_alumni_merge

The second dataset, **active_alumni_merge** is created to investigate the difference between alumni who are out of the legacy pipeline with those who are in the pipeline.

Data Merging This dataset is mainly based on the *ActiveAlum* table with two additional columns `subcategory_Y_count` and `sheet_appear_count` as described in section 4.1.1. Since any IDs outside `merged_pipeline` are considered to be non-participants of the legacy pipeline or non-donors of legacy, the column `label` is created to store such information with value 0 for non-participants and value 1 for pipeline participants.

Sanity Check & Feature Engineering Since **active_alumni_merge** and **merged_pipeline** contain coincident columns, feature engineering for the second dataset mainly follows similar logic as described in section 4.1.3. Specificly, free-text degrees were mapped into {`Undergraduate`, `Postgraduate`, `Doctorate`, `Diploma/Cert`, `Other`, `Unknown`} using common abbreviations

(e.g., BSc/BA/LLB/BEng/BCom/General Course; MSc/MA/LLM/MBA/MPA/MPhil; PhD/DPhil; Diploma/Certificate/Exec/PGCert).

Label Analysis This dataset is extremely skewed in terms of positive and negative samples, in which only 445 out of 192,157 (0.2%) alumni are in the legacy pipeline. Among the alumni population in this table, pipeline participants are of higher age group and have graduated earlier than non-participants, as demonstrated below.

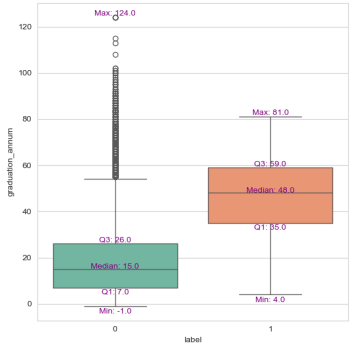


Figure 32: Boxplot for graduation_annum

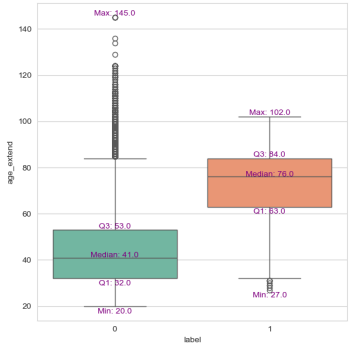


Figure 33: Boxplot for age_extend

The distribution of country among alumni in **active_alumni_merge** is very different from that in **merged_pipeline**. `COUNTRY` of non-participants is dominated by the category Other, followed by the United Kingdom. While, most participants are from the United Kingdom followed by Other. The distribution of `Degree` among participants and non-participants is similar, where Postgraduate and Undergraduate are the top two categories in terms of counts. Because of the huge difference between numbers of participants and non-participants, the following plots are drawn in log scale.

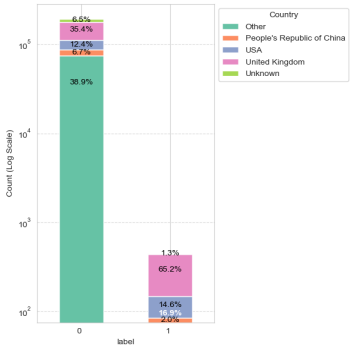


Figure 34: Boxplot for COUNTRY

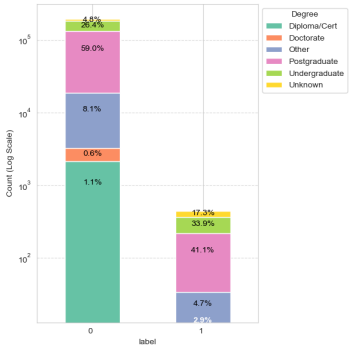


Figure 35: Boxplot for Degree

To avoid vague diagrams caused by skewness of data, the distributions of the rest six variables in label 0 and 1 are presents in tables below. It is notable that the distributions of 2 categories in 2 labels for variables `Marital Status`, `Children Y/N`, and `latest_donation_annum_flag` are opposite, i.e. the dominant categories for label 0 and 1 are different. While, for the other three variables, the trend is the same for 2 labels, which may indicate lack of ability to

distinguish the 2 labels.

Table 4: Distribution of Marital Status

Label	Married	Unknown
0	14231 (7.4%)	177481 (92.6%)
1	166 (37.3%)	279 (62.7%)

Table 5: Distribution of Children Y/N

Label	N	Y
0	186470 (97.3%)	5242 (2.7%)
1	370 (83.1%)	75 (16.9%)

Table 6: Distribution of Alumni Relationship Manager

Label	Unknown	Yes
0	191578 (99.9%)	134 (0.1%)
1	443 (99.6%)	2 (0.4%)

Table 7: Distribution of latest_donation_annum_flag

Label	N	Y
0	176502 (92.1%)	15210 (7.9%)
1	71 (16.0%)	374 (84.0%)

Table 8: Distribution of subcategory_Y_count_flag

Label	N	Y
0	186299 (97.2%)	5413 (2.8%)
1	307 (69.0%)	138 (31.0%)

Table 9: Distribution of sheet_appear_count_flag

Label	N	Y
0	179321 (93.5%)	12391 (6.5%)
1	286 (64.3%)	159 (35.7%)

4.3 Data Summary

A summary of the two datasets is demonstrated below. Before feature engineering, in which most missing values are filled by values following the rules explained previously, 60.1% of cells in **merged_pipeline** and 36.7% in **active_alumni_merge** contain invalid values, such as NAs, empty strings, and the hyphen symbols. The large amount of missing information tends to increase the difficulty of analysis and contort the results obtained from models.

Table 10: Created Datasets Overview

Datasets	No. of Rows	No. of Cols	Invalid Before Feature Engineering	Numeric Cols	Categorical Cols	Boolean Cols
merged_pipeline	875	29	60.1%	8	16	4
active_alumni_merge	192157	13	36.7%	3	10	0

5 Modeling

In the modeling part of this project, the two research questions were addressed via various methods. The analysis on potential donors among pipeline participants was conducted using the table **merged_pipeline**, and that on potential participants among active alumni was based on the dataset **active_alumni_merge**. The unsupervised learning method was used for the alumni datasets, including factor analysis and cluster analysis. The aim was to explore the potential influence of the different feature variables on alumni's donation behaviour. Factor analysis extracted the underlying factors that could explain the majority of the variance, thereby identifying the key dimensions that influence the donations. On this basis, cluster analysis divided the individuals to the groups with similar profile respectively, uncovering cohorts like 'High-Intent Mature Donors' and 'General Potential Donors'. From these exploratory analysis, the relationship between the important feature combinations and the donation behavior, as well as the structured clustering information were clearly discovered. Subsequently, the supervised learning method, i.e. predictive model was introduced. Using the alumni data with labels, it predicted the future donation intention of various individuals. This structure achieved the transition from pattern recognition to behavioral prediction, making the models more interpretable and complete, thus the solid data-driven support was provided for customizing cluster-based and precise strategies for legacy cultivation and marketing.

5.1 Factor Analysis

5.1.1 Data Preparation

To prepare data table for factor analysis, variables that are non-descriptive to characteristics of alumni are excluded, such as `PROPOSAL_TYPE_ECHO`, `ID_NUMBER`, and `PROPOSAL_STATUS_ECHO`. Exploratory factor analysis (EFA) expect continuous, numeric data in nature. Therefore, boolean values were converted to 0s and 1s, and categorical variables were one-hot encoded with reference levels removed, which results to 19 variables obtained from **merged_pipeline** and 17 variables obtained from **active_alumni_merge**. Meanwhile, missing values are not allowed in the input data, so missing values in `graduation_annum` and `age_extend` in both datasets were filled by median values according to which group in `legacy_BN` (in **merged_pipeline**) or `label` (in **active_alumni_merge** an ID belonged to. In addition, both datasets were split into training and testing sets in the ratio of 3:7, in which training sets were used in EFA and test sets were used in CFA, in order to evaluate the robustness of proposed factor structure in unseen data.

5.1.2 Factor Analysis: Potential Donors

Before running EFA, a series of tests were applied to processed data to assess if the data is suitable for factor analysis. The Kaiser-Meyer-Olkin Measure (KMO) evaluates the strength of partial correlations among variables, in which both training and test sets achieved scores around 0.78, indicating mild poor data quality for factor analysis. Moreover, the p-values of Bartlett's Test of Sphericity, testing for correlation among variables, were exactly 0 for

both sets, which may suggest that variables were strongly correlated, but was likely because of the existence of many boolean variables. Variance inflation factor (VIF) was also calculated, aiming to identify multicollinearity among variables. `Children_Y/N_Unknown`, `COUNTRY_CODE_Unknown`, and `in_active` yielded the highest VIFs (greater than 20) in both training and test sets, so these variables were removed before EFA.

There exist 2 variables, `age_extend` and `age_group`, in which the former contained numeric age values and the later was the ordered categorical version of the former. Therefore, only one of them was expected to be kept in the model, and the results for using each of them were compared. Two different methods of EFA, principal component analysis (PCA) and principal axis factoring (PAF), with two rotation functions, Varimax and Promax, were attempted on the training dataset. To start with, EFA was conducted using default settings to produce the scree plots below. According to the elbow points, both models using `age_extend` and `age_group` should choose 3 factors, while Kaiser's Criterion (eigenvalues greater than 1) suggested that 5 factors would be used for both models. Hence, 8 EFA models, of 2 methods, 2 rotations, and 2 sets of data, were firstly fitted using 5 factors.

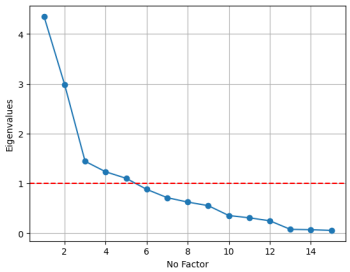


Figure 36: Scree Plot for EFA with age_extend

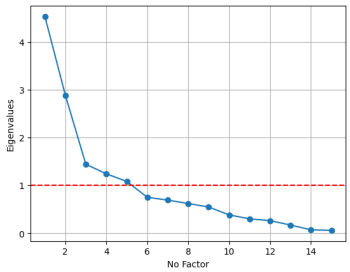


Figure 37: Scree Plot for EFA with age_group

Significant loadings above 0.4 for factors in these 8 models are demonstrated below. It could be noticed that loadings obtained using PCA are generally higher than that obtained using PAF, because PCA takes into count of both common and unique variances explained by variables, exaggerating the importance of variables. For models fitted with Varimax rotation, many variables obtained significant loadings on more than one factors, suggesting that correlation among factors should be assumed and Promax might be more suitable. Calculating communities of variables, some features yielded values greater than 1, which was abnormal and probably caused by too many number of factors. The cumulative variances explained 5 factors were 70.68% for models fitted with `age_extend` and 70.89% for `age_group`.

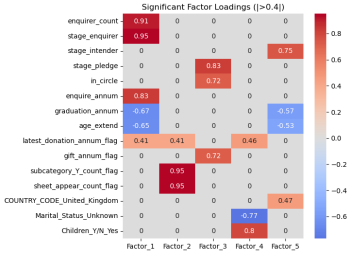


Figure 38: EFA: PCA + varimax + extend

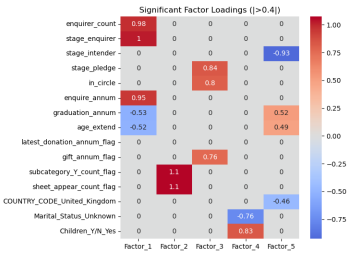


Figure 39: EFA: PCA + promax + extend

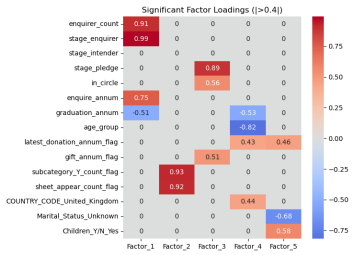


Figure 40: EFA: PAF + varimax + extend

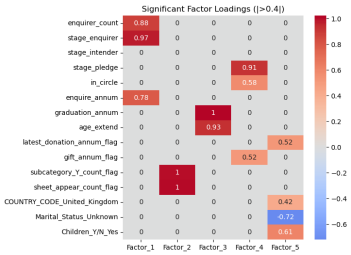


Figure 41: EFA: PAF + promax + extend

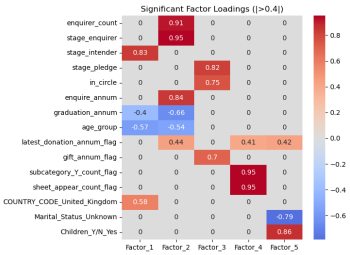


Figure 42: EFA: PCA + varimax + group

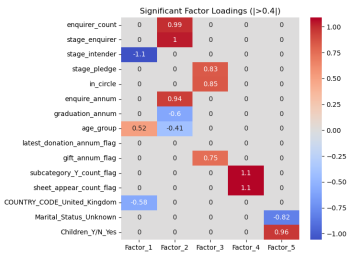


Figure 43: EFA: PCA + promax + group

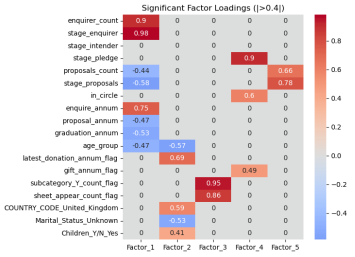


Figure 44: EFA: PAF + varimax + group

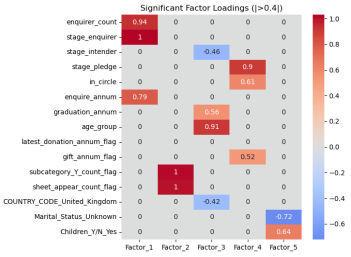


Figure 45: EFA: PAF + promax + group

To improve results of EFA and address issues described above, models using PAF and Promax rotation were further improved by combining and deleting input features. Various modifications were attempted and compared according to loadings, communities, cumulative variance, and determinacy coefficients of factors. Each time a new set of features was created, scree plot was also generated to help decide optimal number of factors. Eventually, the best-performancing models with `age_extend` and `age_group` were obtained using the same set of processed features:

- **gift_annum_flag** and **latest_donation_annum_flag** were summed to create **previous_donation**;
- **stage_intender**, **stage_pledge**, and **in_circle** were joined to generate **pipeline**;
- original variables mentioned above and **graduation_annum**, **Marital_Status_Unknown**, as well as **Children_Y/N_Yes** were then removed.

The dominant variables for factors were generally the same for models using `age_extend` and `age_group`. As shown below, the second factors in both models were strongly, positively correlated to `subcategory_Y_count_flag` and `sheet_appear_count_flag`. However, the left plot below suggests that `age_extend` is correlated to the first factor with loading -0.56, while the right plot shows that `age_group` contributes more to the third factor with loading -0.6. Another interesting point is that the signs of loadings for features highly correlated to the first factors (except `age_extend` are completely opposite for the two models.

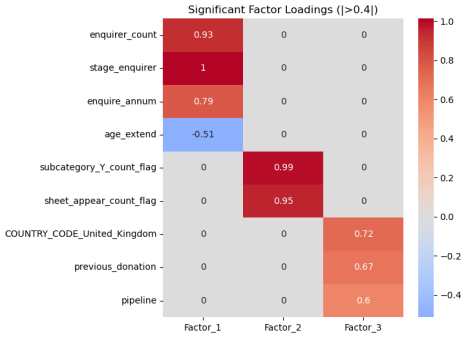


Figure 46: Reduced EFA with age_extend

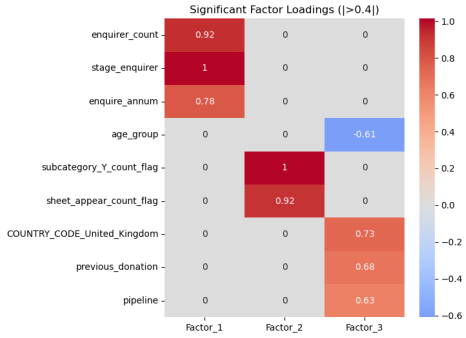


Figure 47: Reduced EFA with age_group

The two EFA models generated similar cumulative explained variances and determinancy coefficients for the first 2 factors, as shown below. However, the determinancy coefficient of the third factor in the model with `age_extend` was only half of that in the model with `age_group`. Therefore, the final model was chosen to be the one with `age_group`.

Table 11: EFA Results on Potential Donors

Evaluation	**Model with age_extend**	**Model with age_group**
Factor 1 Determinacy Coefficient	0.654	0.630
Factor 2 Determinacy Coefficient	0.484	0.476
Factor 3 Determinacy Coefficient	0.268	0.434
Cumulative Explained Variance	77.51	79.77

The proposed factor structure in **merged_pipeline** from the final model was then tested using CFA with maximum likelihood method for estimating parameters. Evaluation metrics, as described in section 3.1.2, and their interpretation are demonstrated below. Although the final model outperformed any other factor structures attempted in EFA, it was still underfits. This is likely because that factor 2 had only two binary indicators, and factor 3 mixed heterogeneous content and includes weak items (low communalities). In addition, several variables were binary/zero-inflated or highly skewed counts, and maximum likelihood estimation with continuous-normal assumptions could inflate misfit.

Table 12: CFA Results on Potential Donors

Evaluation	**Value**	**Interpretation**
Chi-square/df	6.142	substantial misfit relative to df
CFI	0.915	borderline; acceptable incremental fit
RMSEA	0.14	very high; indicates serious misspecification

Local Summary According to the final model, the first factor could be summarised as **Enquiry Engagement**, which is positively correlated to `enquirer_count` , `stage_enquirer` and `enquire_annum`. The second factor is concerned with **Alumni Relation** because it

is highly dependent on two features measuring the connection between alumni with the School. The third factor tends to describe **Personal Profile** of alumni, in which age, country of origin, previous donation records, and participation of pipeline provide information for it. Overall, the top three variables with high communities and high loadings are `sheet_appear_count_flag` (communities 0.96), `stage_enquirer` (communities 0.93), and `subcategory_Y_count_flag` (communities 0.84).

5.1.3 Factor Analysis: Potential Participants

Similar tests were conducted before EFA as described in section 5.1.2. The KMO values for training and test sets were both 0.353, which indicates that the data is not suitable for factor analyaia, and the p-values of Bartlett's Test were both 0. `age_extend`, `graduation_annum` and `Degree_Postgraduate` had VIFs above 20. Since `age_extend` and `graduation_annum` highly correlated and `Degree_Postgraduate` was obtained from one-hot encoding, in initial attempts of EFA, only `age_extend` was removed. The Kaiser's Criterion suggested that 8 factors should be used. The loadings of variables in 4 models using 2 methods and 2 rotations are demonstrated below. Generally higher loadings in PCA models can also be observed. In the models using PAF, some variables had communities greater than 1 and some factors had only 1 dominant variable, which indicates that the number of factor was too large. Since most features only had significant loadings in 1 factor, variables were not highly correlated and using Varimax rotation would be enough.

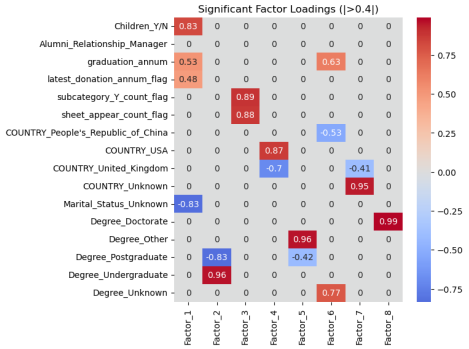


Figure 48: EFA: PCA + varimax

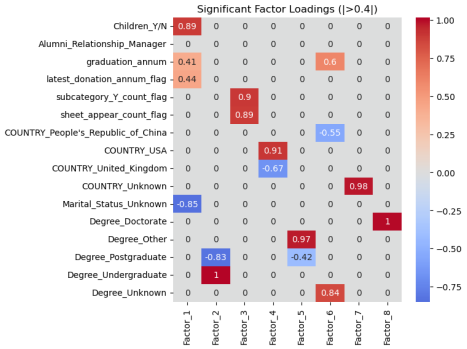


Figure 49: EFA: PCA + promax

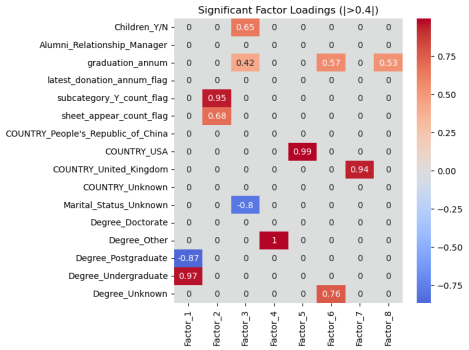


Figure 50: EFA: PAF + varimax

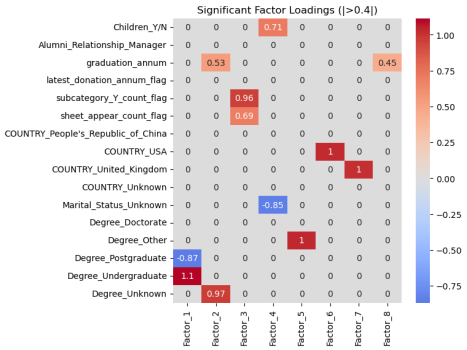


Figure 51: EFA: PAF + promax

The model using PAF and Varimax rotation was improved by integrating `Degree_Undergraduate`

and `Degree_Postgraduate` into `Degree_post_under` and joining `Degree_Doctorate`, `Degree_Other`, and `Degree_Unknown` into `Degree_new_other`. Country-related variables, `Alumni_Relationship_Manager`, as well as, `latest_donation_annum_flag` were also removed. Kaiser's Criterion suggested that this new data table should have 3 factors. The loadings of the final model are shown below. The cumulative variance explained by these 3 factors was 78.61%. The determinacy coefficients for all factors were very low, (0.483, 0.286, and 0.279 respectively for factors 1 to 3), indicating unstable data structure.

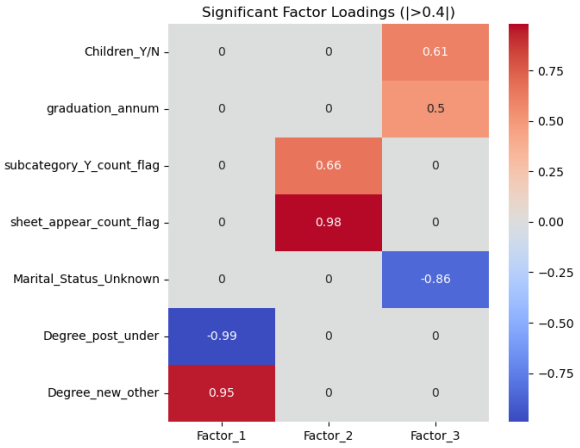


Figure 52: Reduced EFA for Potential Participants

Based on this final model, a similar analysis with CFA as in section 5.1.2 was conducted to assess the robustness of the proposed factor structure. As demonstrated in the table below, the model showed excellent incremental fit, but the Chi-square/degree of freedom was astonishingly high, suggesting misfit of model. RMSEA also suggests existence of local misfit. However, this result should be surprising if considering the data structure in **active_alumni_merge** and poor results in EFA.

Table 13: CFA Results on Potential Participants

Evaluation	**Value**	**Interpretation**
Chi-square/df	397.059	extremely high, exact-fit test rejected
CFI	0.0.979	strong incremental fit relative to baseline
RMSEA	0.083	borderline; suggests some residual misfit

Local Summary The first factor generated by the final model is related to **Degree**, in which the 2 associated variables are dummy variables from the same categorical feature. The second factor is also dominated by 2 features concerned with **Alumni Relation**, but the loading of `subcategory_Y_count_flag` is not as high as in the model for potential donors. The third factor is also similar to the one in previous analysis, in which **Personal Profile** features mainly contribute to this factor. The three variables with the highest loadings and communities all above 0.9 are `Degree_post_under`, `sheet_appear_count_flag`, and `Degree_new_other`.

5.2 Cluster Analysis

5.2.1 Data Preparation and Standardization

The analytical process commenced with a foundational data preparation stage, essential for ensuring the stability and integrity of the subsequent clustering models. The primary objective was to extend past simple demographic segmentation to identify behaviorally distinct personas. The features included had been standardized. This procedure had transformed the data to have a mean of zero and a standard deviation of one, which had been crucial for distance-based algorithms like K-Means. It had prevented features with larger scales (like age) from disproportionately influencing the clustering outcome over features with smaller scales (like pledge status), ensuring that each variable had contributed based on its variance, not its magnitude.

5.2.2 Cluster Analysis for Potential Donors

K-Means Clustering With a standardized dataset, the analysis began with K-Means, a centroid-based partitioning algorithm. A central challenge in using K-Means had been that the optimal number of clusters, 'k', had to be determined beforehand. To address this scientifically, a systematic hyperparameter tuning process conducted. This involved iterating through a range of potential 'k' values and evaluating each outcome using multiple metrics. The "Elbow Method" was used to plot the sum of squared distances from each point to its assigned center (inertia), which helped identify the point where adding more clusters yielded diminishing returns. At the same time, the Silhouette Score was computed for each 'k'. This measure compared how similar a point is to its own cluster and also to others.

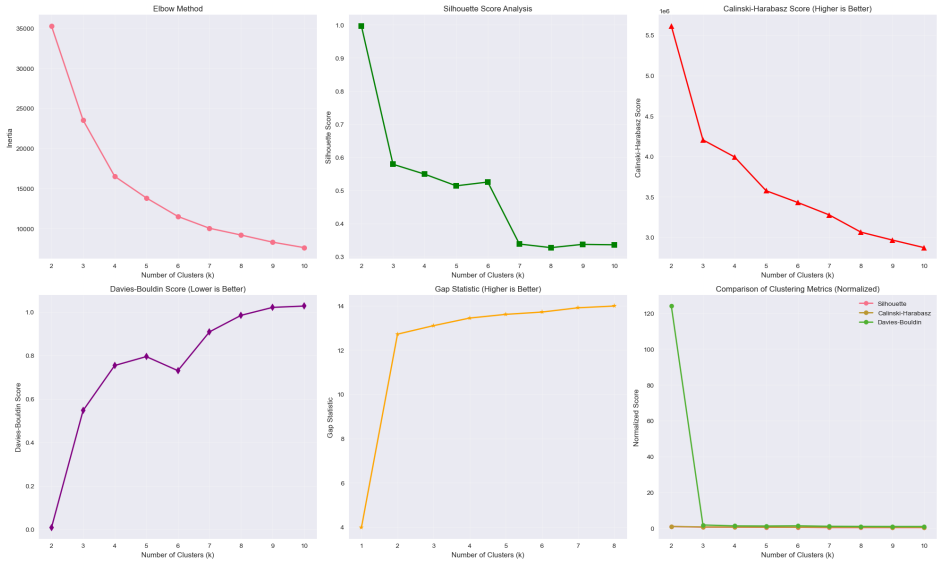


Figure 53: Comparison of Clustering Evaluation Metrics for Optimal k Determination

From Figure 53, choosing k=3 as the cluster number was optimal. In the "Elbow Method", k=3 was the inflection point which the inertia dropped evidently, indicating that adding new cluster numbers would only bring a limitation of effectiveness. Besides, k=3 got a well mark among Silhouette Score ,Calinski-Harabasz Score and Davies-Bouldin , showing that the

structure of clustering is clear. Gap Statistic was steady after reaching a high point at k=3. To conclude, k=3 reached a well balance between stability, structure and interpretability, which was a best choice in this cluster analysis.

Hierarchical Clustering Hierarchical Clustering was employed for further analysis. This procedure was not based on pre-equating the number of clusters, but on the construction of a cluster hierarchy. Various linkage methods were tried out: the criterion to join various clusters, and found that the 'ward' method, which minimizes the total variance in the clusters, gave the most coherent results. As Figure 54 the dendrogram then revealed that the longest vertical distance not crossed by a horizontal line was at the top, suggesting that two statistically strongest (primary) branches emerged.

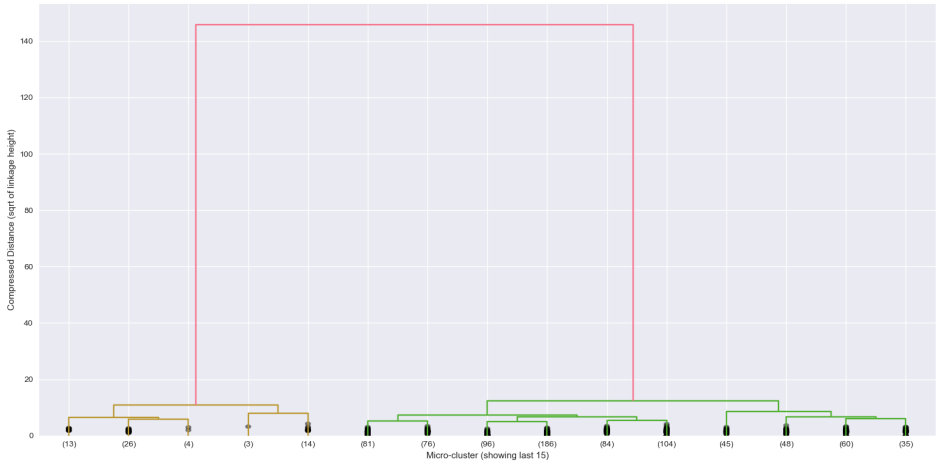


Figure 54: Hierarchical Clustering Dendrogram(Ward linkage)

Exploring Alternative Methodologies Two further models were tested to confirm a comprehensive evaluation. The BGM model, which introduced a probabilistic way, gave a probability for each data point to belong to each cluster. This "soft clustering" might have been useful for data with less-clear batch effects, but the BGM model underperformed in this case. It had a slight tendency to over-segment the data, that is to say, it proposed a structure which was more complex than there really was. This suggested that the underlying clusters were sufficiently separate that the soft, probabilistic assignments were less suitable.

DBSCAN clustering was also applied, which grouped close points together. Although the method of the model was successful to some extent in recognizing the main dense group, it classified a large part of the data as "noise." It proved a devastating error for a fundraising application where the explicit aim was to divide the whole population. Each one of them was a candidate for a relationship, and so it was not politically worthwhile to declare some of them to be analytically irrelevant.

Comparison of Four Clustering Algorithm According to the estimation of four clustering method(K-Means, Hierarchical Clustering, Bayesian Gaussian Mixture Model, DBSCAN) using three internal validation metrics (*Silhouette Score*, *Calinski-Harabasz Index*, and *Davies-Bouldin Index*), Hierarchical Clustering performed the best among all. This method achieves

Table 14: Clustering Algorithm Performance Comparison on Pipeline Dataset

Algorithm	n_clusters	Silhouette	Calinski-Harabasz	Davies-Bouldin	Noise Ratio
K-Means	3	0.5790	4,203,195.2630	0.5465	–
BGMM	8	0.2399	1,977,036.0943	1.5898	–
DBSCAN	10	0.9066	12,311.3268	0.1425	0.7394
Hierarchical	2	0.9964	5,611,905.1029	0.0081	–

the highest *Silhouette Score (0.9964)*, showing that the clusters are well-separated and dense. Meanwhile, its *Calinski-Harabasz Index* reached exceptionally high, suggesting the variance between clusters is much greater than the variance within clusters in Hierarchical Clustering. Furthermore, its *Davies-Bouldin Index* was the lowest, indicating that the clusters generated by Hierarchical Clustering was not only dense, but also well-separated. These indicators illustrated that Hierarchical Clustering could best capture the potential structure of the legacy donor data and could be the most interpretable method for the future profile analysis and strategy making.

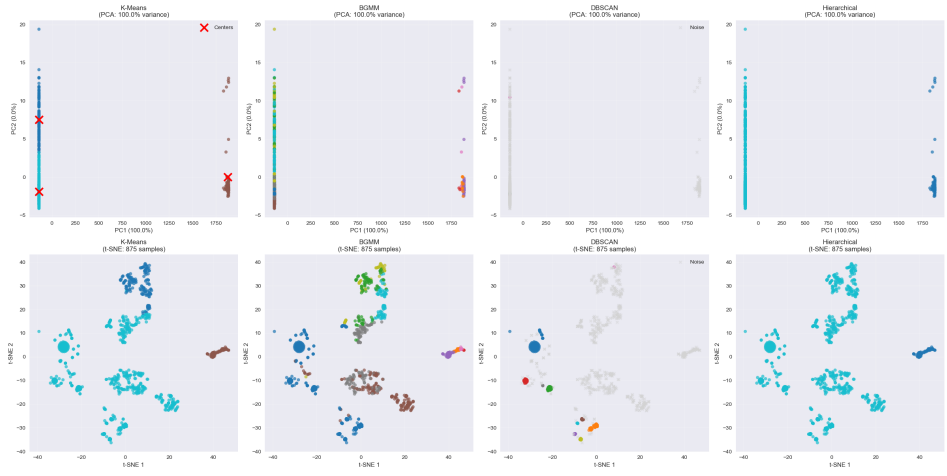


Figure 55: Cluster Result Visualization: t-SNE and PCA

From Figure 55, four clustering method using PCA and t-SNE were visualized to show their performance. In Hierarchical Clustering, the two cluster had a very distinctive separation and clear boundary with well density within cluster, showing a good quality of clustering intuitively. In comparison, although K-means had an overall clear structure, the boundary between clusters overlapped. BGMM had more small clusters, which overlapped with one another, indicating excessive clustering. DBSCAN had many noise point distributing at the edge with a sparse structure. This figure visually verified the rationality of choosing Hierarchical Clustering as the best algorithm.

In this cluster analysis for merged_pipeline dataset, Ward's minimum variance method was used for Hierarchical Clustering. The distance defined in this method is not direct Euclidean distance but the augumenter of Sum of Squared Errors after combining two clusters. This definition is rational because it tends to prioritize the merging of clusters where variance increase is the smallest after the merge, achieving a more compact and internally similar clustering structure.

Results and Interpretation The analysis had revealed a definitive binary structure within the alumni population. The rigorous evaluation of four distinct clustering models had provided convergent evidence that the most powerful differentiating factor had been the act of making a legacy pledge. The fact that Hierarchical Clustering had arrived at this conclusion with near-perfect Silhouette Scores (0.9964) had confirmed this structure with exceptional statistical confidence. It had revealed a powerful binary of those who "had-pledged" and those who "had-not," providing an unambiguous foundation for strategic planning.

Table 15: Cluster Feature Comparison Summary

Feature	**Cluster 1**	**Cluster 2**	**Overall Average**
AGE	80.28	70.11	71.40
proposal_annum	9.87	7.99	8.12
stage_pledge	1.00	0.16	0.22
enquire_annum	1.33	1.90	1.86
enquirer_count	0.18	0.31	0.31
stage_enquirer	0.17	0.29	0.28
stage_intender	0.12	0.14	0.14
proposals_count	0.95	0.84	0.84
graduation_annum	54.37	49.34	49.76
Cluster Size (%)	60 (6.9%)	815 (93.1%)	875 (100%)

Table 15 showed the feature comparison between two clusters. The first five features were the relatively more important ones based on their high variance across clusters.

Cluster Profile 1: The High-Value Legacy Donors

The first group, termed the "High-Value Legacy Donors", had been a small but extremely high-value segment, making up just under 7% of the population. This group's had two key features: the presence of a `stage_pledge` in the dataset, indicating a commitment of giving, and a significantly older demographic, with an average age of approximately eighty years. This cluster had represented the bedrock of the legacy program: the confirmed, loyal, and mature donor base already deeply committed to the institution's future. They had been the guardians of the philanthropic mission, and the strategic focus for this group should have been on high-touch stewardship and recognition to reinforce their vital commitment.

Cluster Profile 2: The General Alumni Population

In stark contrast, the second group, the "General Alumni Population", had comprised the vast majority (over 93%) of individuals. The `stage_pledge` value for this cluster had been very low. This cohort had been a full decade younger than the pledged donors, with an average age of seventy. This cluster represented a vast reservoir of untapped potential. While they had been uncommitted at the time, they had been the next generation of philanthropists. The strategic imperative for this group had been long-term cultivation, involving broad educational campaigns about legacy giving, followed by targeted, data-driven messaging to nurture interest and guide them on their journey toward making their own commitment.

In this cluster analysis, the potential legacy donors had been divided to two notably different groups. Cluster 1 had high donation conversion rate with more serious commitment, especially very active in Pledge Stage and final Proposal Stage. Their age were also remark-

ably higher than the average. According to their behavior characteristics, this group could be defined as 'High Value Segment' and high maturity participants. The recommendation was to maintain continuous and thorough follow-up and focus on promoting loyalty and donation fulfillment rate. Cluster 2 had a very large population,with a high level of participation in Enquirer Stage, showing widespread but superficial interactions. This group was defined as 'Prospect Expansion Segment', having significant room for growth in donation intention and formal pledge. The strategy could be enhancing the motivation and cultivation technique by directional communication, emotional connection and donation value guidance to push them converting to a further stage.

5.2.3 Cluster Analysis for Potential Participants

In this section, the object of study was all active alumni with a population of 192157. The methodology used is similar with the previous section, but concerning a much larger dataset, the result to be explored was of wider significance.

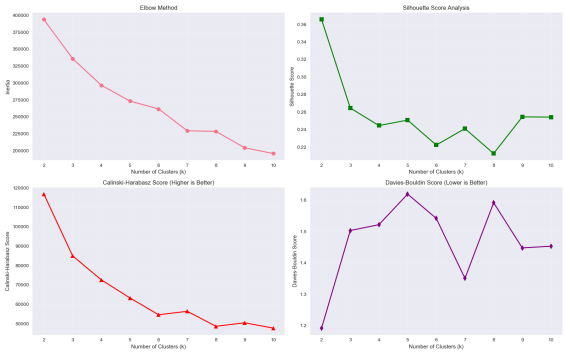


Figure 56: Comparison of Clustering Evaluation Metrics for Optimal k

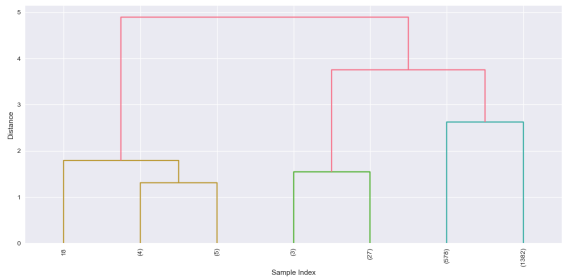


Figure 57: Hierarchical Clustering Dendrogram

K-Means Clustering Concerning K-means clustering method, the optimal k chosen this time was different. As the Figure 56 showed, in Silhouette Score, Calinski-Harabasz Score and Davies-Bouldin Score, k=2 all performed the best, indicating well density of clusters and good variance between clusters. Also, the overlapping of the clusters were the least. In the elbow method, the inertia did not decrease heavily as the k grew, so the marginal benefit of adding new clusters was not much. The recommended k for K-means in active alumni case was 2.

Hierarchical Clustering In Hierarchical Clustering, the problem of the large population dataset existed as computational cost was huge for this algorithm. To solve this trouble, PCA and small batch of samples (the quantity of 2000) were used to evaluate different linkage methods such as ward, average and complete under different cluster number k. The best linkage and k was chosen by calculating the best silhouette score. Subsequently, to avoid memory overflow caused by large size of the distance matrix on the whole dataset, k-nearest neighbors graph was applied, effectively reducing the computational complexity. This strategy maintained the effect of model as well as considered the limit of computational

limitation, letting the Hierarchical clustering apply successfully on the whole dataset. In the experiment, the average linkage and k=2 was chosen for the algorithm, which is consistent with K-means result in the choice of the cluster number.

Exploring Alternative Methodologies The BGM model had a high demand on memory and computing on a large scale dataset, so small batch samples had been trained to obtain the number of active components, then the whole dataset were predicted. To raise efficiency, lighter covariance matrix type was used for high dimensions, making the clustering process extensible. In the BGMM case, the optimal k was 9.

With regard to DBSCAN, this method also had a high computation cost on large dataset, which may cause memory corruption, A subset of 10000 samples was trained to adjust parameter, choosing the best epsilon value and applying on the whole dataset. Beside, the data type was transfered to float32 while the spatial index structure was set to ball_tree, accelerating the process. This measurement significantly reduced computational resource consumption and also maintain the robustness of the result k=4.

Table 16: Clustering Performance Comparison on Active Alumni Dataset

Algorithm	**n_clusters**	**Silhouette**	**Calinski-Harabasz**	**Davies-Bouldin**	**Noise Ratio**
K-Means	2	0.3656	116814.8801	1.1919	–
DBSCAN	4	0.3081	111.5278	1.1179	0.0003
Hierarchical	2	0.1624	30018.5989	2.1513	–
BGMM	9	0.2215	29662.7471	1.8004	–

Comparison of Four Clustering Algorithm For the comparison of the cluster analysis on the active alumni dataset, the four methods above were evaluated. K-Means had the outstanding *silhouette score*, showing the high density within cluster and good separation between clusters. Besides, K-Means also had the highest *Calinski-Harabasz index* and relatively low *Davies-Bouldin index*, verifying its good quality of clustering. To conclude, K-Means won the best among all clustering methods.

K-Means cluster algorithm was chosen for active_alumni_merge dataset, whose distance was defined as Euclidean distance. After the standardization of features, Euclidean distance can make sure each dimension has the same weight in the similarity calculation, so as to fairly measure the geometric proximity degree between the sample and the cluster center. This measurement is consistent with the objective function to minimize Within-Cluster Sum of Squares, ensuring the rationality and interpretability of the cluster results.

Results and Interpretation After applying the final best algorithm K-means, the whole population is divided into two separate groups. As Figure 58 shows, Cluster A had a population of 129292, while Cluster B had a population of 62865. The clustering quality was better evaluated not only through internal validation metrics (e.g., *Silhouette*, *Calinski-Harabas*) but also through feature profile differences. The analysis confirmed KMeans achieved meaningful segmentation based on feature distribution.

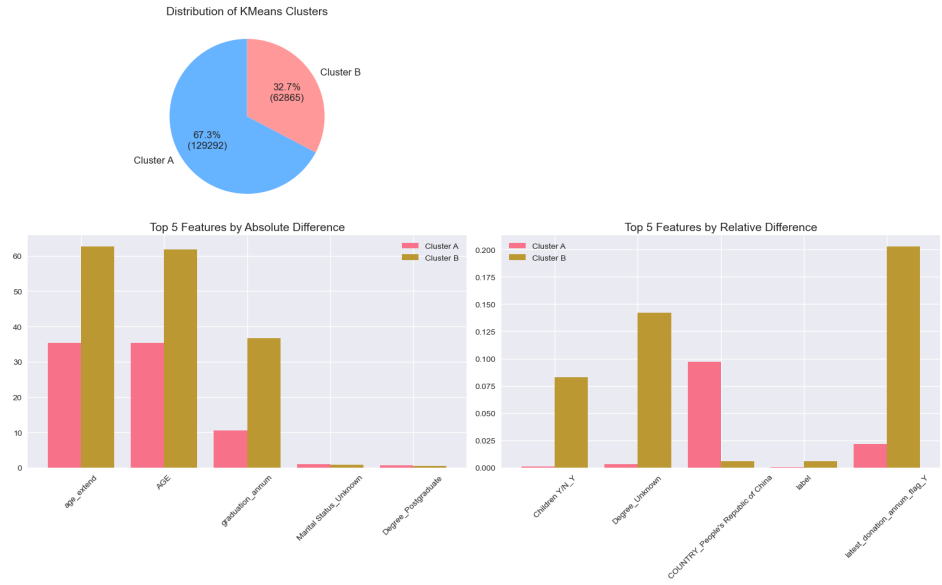


Figure 58: Distribution of K-Means Clusters and Top Features

Cluster Profile A: General Potential Donors

The first group making up 63.7% of the population could be regarded as 'General Potential Donors'. The main characteristic of this group was the young age as the average age of the members was only about 35 years old. Their time after graduation was merely around 10 years, showing they were still in the early stage of their life. This cohort were more likely to be Unknown in marriage and degree, indicating the tendency of their information incompleteness. This cluster barely not had the history record of last donation as well as pipeline joining, so they had not been formed a steady donation behavior or clear donation intention yet.

Although the conversion rate of them was relatively low, this group was the largest component of the active the alumni. As they were still in a young age, they could become the long-term cultivating candidates in the future. The suggestion would be maintaining continuous communication and following up to improve their information network. Also, face-to-face alumni association and online platform interaction should be enhanced to strengthen their emotion bond with the alma mater. In these ways, their agreement of donation and the willingness to participate would be gradually cultivated.

Cluster Profile B: High-Intent Mature Donors

The second group accounting for 32.7% of the total active alumni displayed the characteristic of 'High-Intent Mature Donor' despite its smaller population. The average age of this group is above 61 years old, much elder than the cluster A. Meanwhile, the members' year after graduation was three times higher than the first group, showing that they were not only older but also owned a longer time of alumni position. Besides, this group showed a stronger donation intention in terms of behavior such as they were much more likely to have a donation history and to join the legacy pipeline, indicating their more explicit expression of donation aspiration. This cohort also had a higher proportion of members who had a child or more, probably because they were in a further stage of life than the first group.

This group was smaller in size but more core, needing more elaborate cultivating strategies such as establishing a membership system for these key alumni, with appropriate honors bestowed upon them, so as to enhance their sense of belonging. Besides, to adopt a customized operation on them was a good choice. The team could invite members to attend the legacy donation information session and push personalized donation stories and examples, which could also strengthen their donation willingness as well as guide them to officially enter the final legacy giving process.

5.3 Predictive Models for Potential Donors

Dataset and Feature Dataset 4.1, with a filter of `is_pipeline`=1, is used to predict whether an individual will make a legacy donation, which it contains 479 rows of data with 290 negative and 189 positive.

The binary variable `label` in the dataset indicates whether an individual has already made a legacy donation, where it means the individual appears in the `proposal` table when label=1, and vise versa.

The dataset is split into training and test sets at a 1:1 ratio. 15 features were selected that are informative for modeling and do not leak label information such as features from dataset *Proposal*. The baseline models follow those listed in 3.3; each model uses a fixed set of hyperparameters. When different models require the same hyperparameter, the same value was assigned across those models. For tree-based models, we uniformly set $n\_estimators = 25$. Given the dataset size, preliminary experiments indicate this setting is sufficient to achieve solid baseline performance, although it looks relatively small.

Model Evaluation Metrics For each model, the following evaluation metrics were calculated on the test set: **accuracy, precision, recall, f1-score, and AUC**. Additionally, the **ROC curves** for all models are plotted to compare their performance. Further analysis was performed on each model's output coefficients and feature importance scores to identify the most influential predictors. The definitions of metrics used to evaluate model performance:

- **Accuracy**: The proportion of correct predictions among all predictions.
- **Precision**: The proportion of correctly predicted positive cases out of all predicted positives.
- **Recall**: The proportion of actual positive cases that were correctly identified by the model.
- **F1-score**: The harmonic mean of precision and recall.
- **AUC**: The area under the ROC curve, measuring the model's ranking capability.
- **ROC Curve**: A plot of the true positive rate versus the false positive rate at various classification thresholds.

Hyperparameter Tuning and Models Stacking For the fine-tuned models, two algorithms- *grid search* and *random search*- were compared over the same hyperparameter spaces to assess both the computing consumption and models performances. Both algorithms used 5-fold

cross-validation with ROC-AUC as the scoring metric. Random search samples $n\_iter = 30$ configurations with a fixed seed for reproducibility.

Table 17: Models and Hyperparameter Spaces (Concise Rationale)

Model	**Hyperparameter**	**Search Space**	**Rationale**
Random Forest	`n_estimators`	[20, 25, 30]	Trees; bias–variance, stability.
	`max_depth`	[`None`, 2, 3, 5, 8, 10]	Limit depth; prevent overfit.
	`min_samples_split`	[2, 3, 5]	Avoid tiny noisy splits.
	`min_samples_leaf`	[1, 2, 4]	Smoother leaves.
	`max_features`	[`sqrt`, `log2`]	Split diversity.
	`bootstrap`	[`True`, `False`]	Variance reduction.
GBDT	`n_estimators`	[20, 25, 30]	Boosting stages.
	`learning_rate`	[0.02, 0.05, 0.1]	Shrinkage; generalization.
	`max_depth`	[2, 3, 5]	Interaction order; overfit control.
	`subsample`	[0.6, 0.8, 1.0]	Stochastic rows; variance↓.
XGBoost	`n_estimators`	[20, 25, 30]	Boosting rounds.
	`learning_rate`	[0.02, 0.05, 0.1]	Shrinkage step.
	`max_depth`	[2, 3, 5]	Max tree depth.
	`subsample`	[0.6, 0.8, 1.0]	Row subsampling.
	`colsample_bytree`	[0.6, 0.8, 1.0]	Column subsampling.
Logistic Regression (L1)	`C`	[0.01, 0.05, 0.1, 1, 10]	Inverse reg.; smaller ⇒ sparser.
	`penalty`	[`l1`]	L1 sparsity.
	`solver`	[`liblinear`]	Supports L1.
	`max_iter`	[1000]	Ensure convergence.

Search algorithms **Grid Search** exhaustively evaluates the Cartesian product of all hyperparameter values in the specified grid, guaranteeing the best score within that grid, but at a higher computational cost. While **Random Search** samples a fixed number of configurations from the same grid (treated as discrete distributions here), often discovering near-optimal settings with far less time. The hyperparameter space settings are shown in Table 17

Optimal Model Selection The *optimal model* is the one that achieves the best overall performance under evaluation metrics of 5.3, then its feature importance will be shown and explained, also it will be used to predict possibility and labels to those who have not committed to legacy donation.

5.3.1 Baseline Models Results

The baseline models are trained on some fixed selected hyperparameters. Models that share the same hyperparameters are set at the same value for fair comparisons.

Among the three models based on Logistic Regression, the regular one reaches the highest on AUC (91.55%), with high and balanced results among other metrics. The Bagging model didn't win over the regular one.

Based on the tree models, GBDT, XGBoost, and Random Forest are the top-3 on AUC. GBDT are also relatively higher on other metrics as well, making it a good candidate for further fine-tuning. XGboost also shows its stability in predicting, comes to the second place in all tree models on different metrics. The bagging tree model didn't perform better than boost tree models, nearly the same as Random Forest, therefore fine-tuning work will be performed on Random Forest only instead of Bagging Tree. Decision Tree gets poor results in the default settings, further fine-tuned work on it is not considered. As a result, Random Forest, GBDT, XGBoost, and Logisitc Regression (l1) are selected to be fine-tuned later among all the models.

Key Findings GBDT attains the highest *AUC* (92.72%). A narrow second cluster follows: Logistic Regression (91.55%), Bagging (Logistic) (91.54%), and XGBoost (91.52%), then L1 (91.48%). The top *F1-score* (label=1) is achieved by Logistic Regression (82.72%), with L1 (82.29%), GBDT (82.16%), and Bagging (Logistic) (82.11%) close behind. The best *Recall* (label=1) is a tie between Logistic Regression and L1 (both 83.16%), while the best *Precision* (label=1) comes from GBDT (84.44%). The highest *Accuracy* is shared by Logistic Regression and GBDT (both 86.25%). Random Forest and Bagging (Tree) show lower recall ($\approx 71.6\%$), which depresses F1; the single Decision Tree is the weakest overall.

Model selection for fine-tuning For the first binary classification task, model selection prioritised a balance between precision and recall, with F1-score and AUC as key decision metrics due to the imbalanced nature of the dataset. GBDT achieved the highest overall performance with an F1-score of 11.28% and the highest AUC (96.82%), although recall was modest (48.89%). XGBoost delivered the highest recall (62.22%) with competitive AUC (96.15%), indicating strong potential after tuning. Random Forest exhibited a high recall (60.00%) and stable performance across metrics, making it a valuable non-boosting benchmark. Among the linear models, Logistic Regression variants achieved high AUC (≈96.3%) and moderate recall, making them suitable for probability calibration and threshold optimisation. We therefore select **GBDT**, **XGBoost**, **Random Forest**, and **Logistic Regression (L1)** as fine-tuning candidates for this task.

Table 18: Evaluation Results of Different Models (Baseline)

Model	**Accuracy**	**Precision (label=1)**	**Recall (label=1)**	**F1-score (label=1)**	**AUC**
Logistic Regression	86.25%	82.29%	83.16%	82.72%	91.55%
Logistic Regression (L1)	85.83%	81.44%	83.16%	82.29%	91.48%
Decision Tree	79.17%	71.43%	78.95%	75.00%	88.04%
Random Forest	80.00%	76.40%	71.58%	73.91%	89.99%
Gradient Boosting	86.25%	84.44%	80.00%	82.16%	92.72%
XGBoost	83.75%	81.11%	76.84%	78.92%	91.52%
Bagging (Tree)	80.00%	76.40%	71.58%	73.91%	89.79%
Bagging (Logistic)	85.83%	82.11%	82.11%	82.11%	91.54%

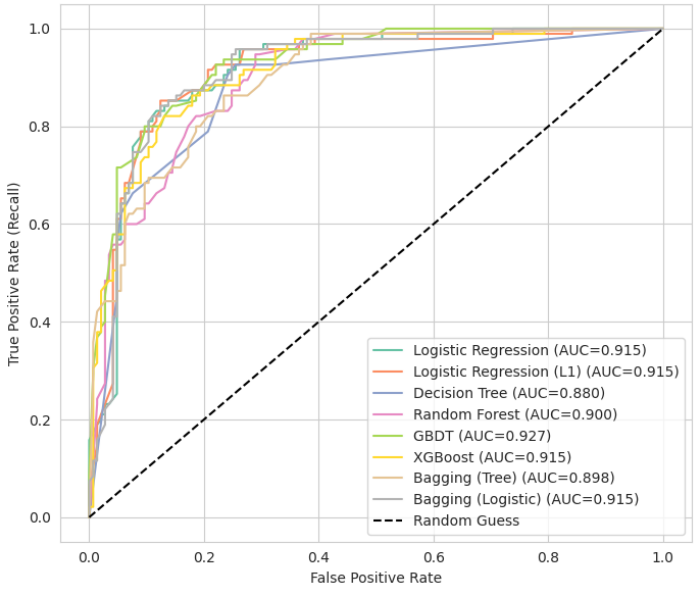


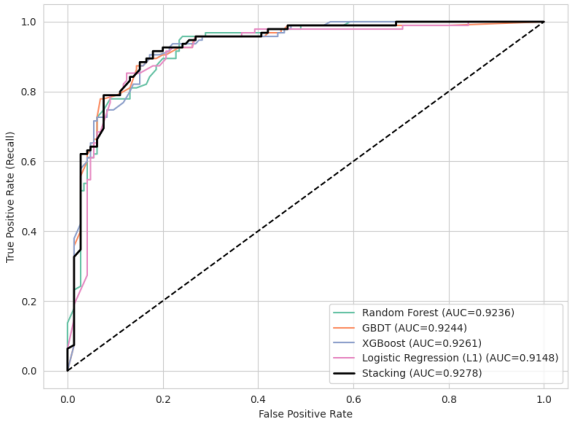
Figure 59: ROC Curves of Baseline models

5.3.2 Fine-tuned and Stacking Models Results

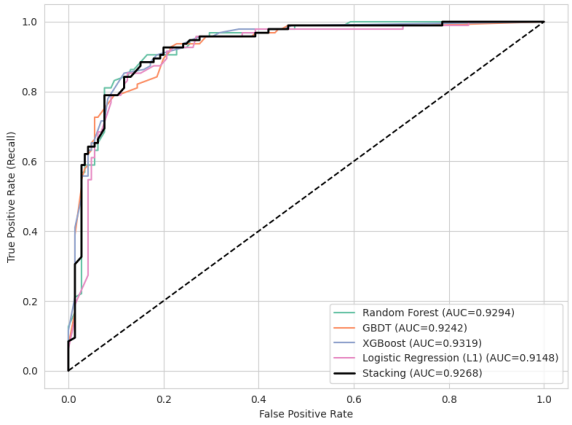
The assessment metrics of fine-tuned models using Grid and Random Search and their stacking models are shown below.

Table 19: Tuned Model Results + Stacking under Grid Search and Random Search

2*Model	Grid Search (4 mins)					Random Search (50 s)				
	Accuracy	Precision (label=1)	Recall (label=1)	F1-score (label=1)	AUC	Accuracy	Precision (label=1)	Recall (label=1)	F1-score (label=1)	AUC
Random Forest	85.00%	83.15%	77.89%	80.43%	92.36%	86.25%	81.63%	84.21%	82.90%	92.94%
GBDT	85.42%	88.46%	72.63%	79.77%	92.44%	85.42%	83.33%	78.95%	81.08%	92.42%
XGBoost	85.00%	85.54%	74.74%	79.78%	92.61%	82.50%	89.55%	63.16%	74.07%	93.19%
Logistic Regression (L1)	85.83%	81.44%	83.16%	82.29%	91.48%	85.83%	81.44%	83.16%	82.29%	91.48%
Stacking Model	85.00%	82.42%	78.95%	80.65%	92.78%	85.83%	82.11%	82.11%	82.11%	92.68%



(a) ROC(GridSearch)



(b) ROC(RandomSearch)

Figure 60: ROC curves for different groups of models

Optimal Model Based on the updated results in Table 19, the **Random Forest** tuned via **Random Search** is selected as the single best model. It achieves the highest F1-score (82.90%) and the highest Recall for the positive class (84.21%), while maintaining competitive Accuracy (86.25%), Precision (81.63%), and AUC (92.94%). This balance between precision and recall indicates the lowest miss rate on positives without materially sacrificing specificity, making it the most suitable choice for deployment among the tuned candidates.

Feature Importance For the selected **Random Forest (Random Search)** model, the most influential feature is `age_group`, contributing 13.55% to the model's predictive power. This is followed by `in_circle_Y` (12.18%) and `enquire_annum` (11.83%), indicating that engagement frequency of related activities are critical in predicting the target outcome. Other notable predictors include `in_circle_N` (8.88%) and `stage_enquirer_False` (8.51%), suggesting that membership circle status and engagement stage have significant explanatory power. Together, the top five features account for over 54% of the model's importance, highlighting a strong concentration of predictive influence in a small subset of variables.

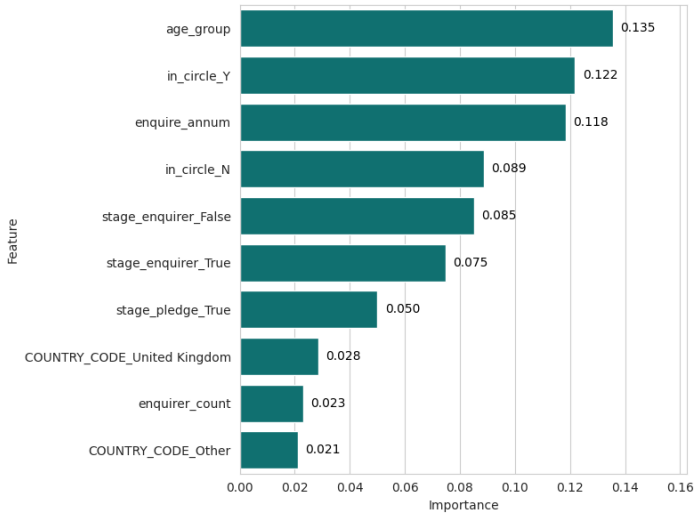


Figure 61: Feature Importance of Random Search Random Forest

5.4 Predictive Models for Potential Participants

Dataset and Feature Dataset 4.2 is used to predict whether an individual will become a participant to join any stage along the pipeline of legacy donation, which it contains 192157 rows of data with 191712 negative and 445 positive, with

The binary variable `label` in the dataset indicates whether an individual has already engaed in the pipeline of making a legacy donation, where it means the individual appears in the merged_pipeline_2 table as well as in the anonymised table when label=1, and one that exists only in the anonymised tablewhen label=0.

Given that the positive class accounts for only 0.23% of the dataset, firstly 10% of the full dataset is set aside as the *final test set* to be used exclusively for the last-stage of evaluation. The remaining data are downsampled prior to training in order to mitigate extreme class imbalance. The ratio for downsampling is set to 1:10 (positive:negative), ensuring that

the training data retain sufficient positive samples for model learning while reducing the dominance of negative samples.

The downsampled dataset is used to train the baseline models, which are then evaluated on the reserved test set using the same metrics framework as in the previous prediction task. Based on this evaluation, candidate models for fine-tuning are selected.

For fine-tuning, the downsampled dataset is further split randomly into 80% training and 20% validation subsets. The same hyperparameter search strategies as before - *Grid Search* and *Random Search* - are applied, using identical parameter spaces to those in the previous subsection.

Finally, the fine-tuned best-performing models are combined into a stacking ensemble, from which the *optimal model* is selected according to the same evaluation criteria as in the previous section. The procedures for feature importance analysis, probability prediction, and labels assignment ($\hat{y} = 1$ if $\hat{p} \geq 0.3$) follow exactly the same steps as described earlier and are therefore omitted here for brevity.

5.4.1 Baseline Models

Key Findings Gradient Boosting achieves the highest *AUC* (96.82%), followed closely by Logistic Regression (96.32%), Bagging (Logistic) (96.35%), Logistic Regression (L1) (96.38%), and XGBoost (96.15%). The top *F1-score* (label=1) is delivered by Gradient Boosting (11.28%), with XGBoost (9.05%) and Bagging (Logistic) (8.60%) as the next best, while Logistic Regression and its L1 variant are close behind (≈8.5%). The best *Recall* (label=1) comes from XGBoost (62.22%) and Random Forest (60.00%), with Bagging (Tree) (57.78%) following; these models also exhibit the highest sensitivity to the positive class. In contrast, Gradient Boosting achieves the highest *Precision* (label=1) at 6.38% but with more modest recall. The best *Accuracy* is obtained by Gradient Boosting (98.20%), while most other models achieve around 97%. The single Decision Tree performs the weakest overall, with the lowest AUC (94.22%) and F1-score (7.69%).

Model selection for fine-tuning For this second binary classification task, the selection criteria prioritised maintaining a high *AUC* while achieving a balanced *precision–recall* trade-off, given the highly imbalanced dataset. Gradient Boosting Decision Tree (GBDT) was chosen for its top AUC (96.82%) and the highest F1-score, despite only moderate recall. XGBoost was selected for its highest recall (62.22%) combined with a competitive AUC (96.15%), indicating strong potential after parameter tuning. Random Forest also demonstrated high recall (60.00%) with a solid AUC (93.87%), making it a valuable non-boosting benchmark. Logistic Regression (L1) was retained for its consistently high AUC (96.38%) and potential for probability calibration. Consequently, the fine-tuning candidates for this task are **GBDT**, **XGBoost**, **Random Forest**, and **Logistic Regression (L1)**.

Table 20: Evaluation between models (Detailed %)

Model	**Accuracy**	**Precision (label=1)**	**Recall (label=1)**	**F1-score (label=1)**	**AUC**
Logistic Regression	97.43%	4.65%	51.11%	8.52%	96.32%
Logistic Regression (L1)	97.39%	4.58%	51.11%	8.41%	96.38%
Decision Tree	97.25%	4.17%	48.89%	7.69%	94.22%
Random Forest	96.50%	3.96%	60.00%	7.44%	93.87%
Gradient Boosting	98.20%	6.38%	48.89%	11.28%	96.82%
XGBoost	97.07%	4.88%	62.22%	9.05%	96.15%
Bagging (Tree)	96.30%	3.62%	57.78%	6.82%	93.44%
Bagging (Logistic)	97.46%	4.69%	51.11%	8.60%	96.35%

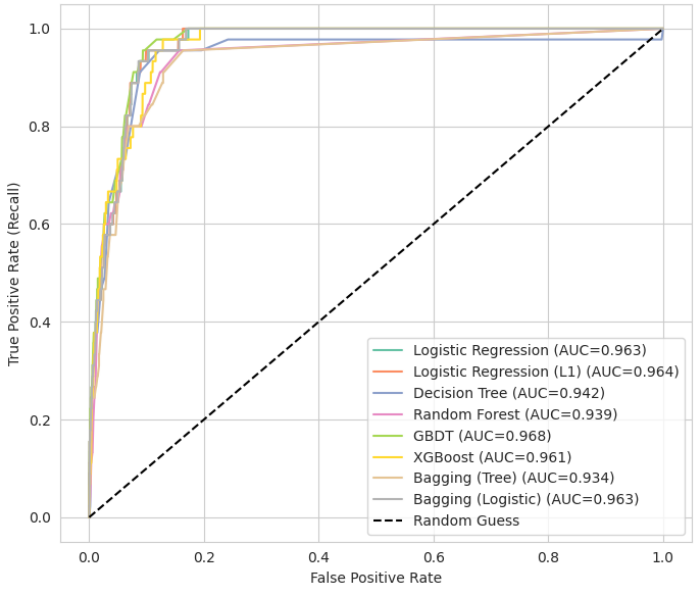


Figure 62: ROC Curves for Baseline Models (on Test Set)

Before proceeding with model fine-tuning, different threshold values were tested for each model to examine the trends of *precision*, *recall*, and *F1-score*, in order to determine whether a better-performing threshold could be identified.

The *precision* and *F1-score* are relatively low across all models, with the highest value reaching only 0.40 for LightGBM, which simultaneously leads to a substantial drop in *recall*. In the context of this project, the primary objective is to increase the amount of legacy donations. Therefore, greater emphasis can be placed on identifying as many potential candidates as possible, making it more desirable to maintain a high level of *recall*. Following this approach, the classification threshold was set to 0.3 to ensure a *recall* of over 60% across all models.

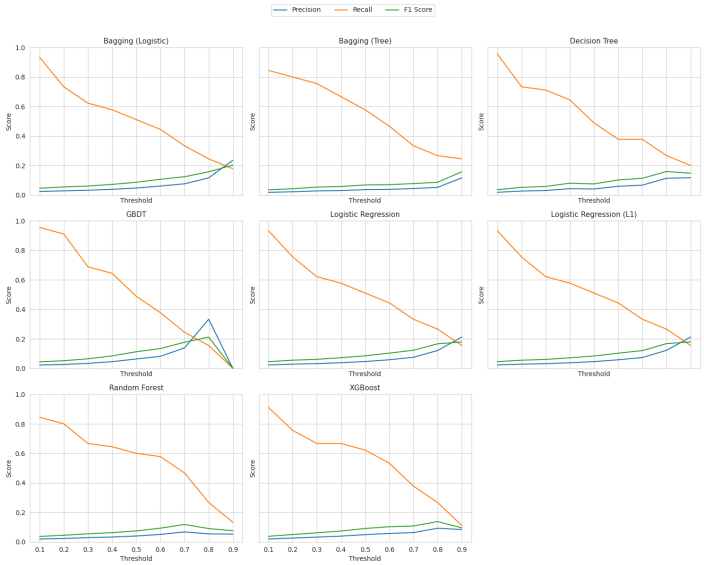
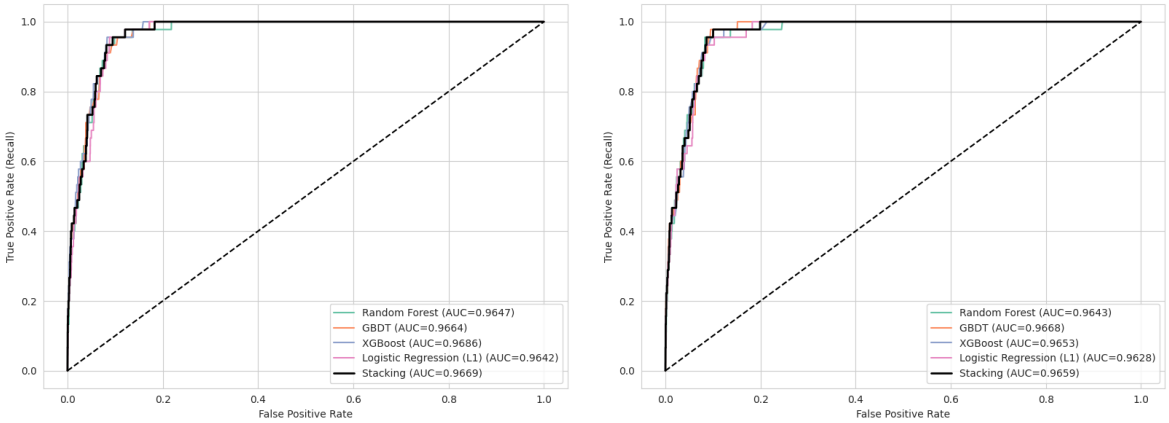


Figure 63: Model Metrics vs Threshold

5.4.2 Fine-tuned and Stacking Models Results

Table 21: Tuned Model Results (Threshold = 0.3) under Grid Search and Random Search

2*Model	Grid Search 7 mins)					Random Search (1 mins)				
	Accuracy	Precision (label=1)	Recall (label=1)	F1-score (label=1)	AUC	Accuracy	Precision (label=1)	Recall (label=1)	F1-score (label=1)	AUC
Random Forest	93.68%	2.90%	80.00%	5.60%	96.47%	94.40%	3.01%	73.33%	5.78%	96.43%
Gradient Boosting	94.81%	3.33%	75.56%	6.38%	96.64%	95.14%	3.16%	66.67%	6.04%	96.68%
XGBoost	94.67%	3.33%	77.78%	6.39%	96.86%	95.27%	3.45%	71.11%	6.58%	96.53%
Logistic Regression (L1)	94.85%	3.08%	68.89%	5.89%	96.42%	95.31%	3.17%	64.44%	6.05%	96.28%
Stacking (LR meta)	96.08%	3.66%	62.22%	6.92%	96.69%	96.46%	3.78%	57.78%	7.10%	96.59%



(a) ROC(GridSearch) (b) ROC(RandomSearch)

Figure 64: ROC curves for different groups of models

Optimal Model Selection Considering both Grid Search and Random Search results, **XGBoost** demonstrates a strong balance between high AUC (96.53%) and competitive precision(3.45%), while delivering a significantly higher recall compared to the Stacking model

(Grid: 71.11% vs 62.22%; Random: 71.11% vs 57.78%). Although the Stacking model attains the highest F1-score by a small margin (Grid: 7.38% vs 6.97%), the recall gain from XGBoost is substantial, making it more suitable for this task, where capturing as many alumnis as possible to be potential candidates for legacy donation is a priority.

Random Forest achieves the highest recall in Grid Search (84.44%), but its precision is notably lower (3.09%), resulting in the lowest F1-score among top models. This indicates that while Random Forest identifies many positive cases, it also generates more false positives, which may not be beneficial enough for the current application.

In summary, **XGBoost** offers the best trade-off considering recall, precision, and AUC, providing strong generalisation performance while prioritising the detection of positive cases. It is therefore selected as the final model for deployment and further analysis.

Feature Importance From the feature importance results of the best XGBoost model under Random Search, the model's predictions rely heavily on a small number of key variables. The feature `latest_donation_annum_flag_N` stands out with a dominance score of 0.7886, far exceeding all others. This indicates that the **absence of a recent donation** is a decisive factor in predicting the target variable, suggesting a very strong association between not having donated in the latest year and the prediction outcome.

The second most important feature, `sheet_appear_count_flag_N` (0.0345), while much less influential than the top feature, still provides supportive information for classification.

Overall, the model's predictive power is highly concentrated on the `latest_donation_annum_flag_N`, with all other features contributing much less. This high concentration implies that, on one hand, the model captures a strong single-variable signal; on the other hand, there is a potential **risk of over-reliance on one feature**. That said if this variable's data quality or distribution changes, the model's generalization ability could drop significantly. Therefore, it would be advisable in future work to incorporate more high-quality, complementary features to improve the model's stability and robustness.

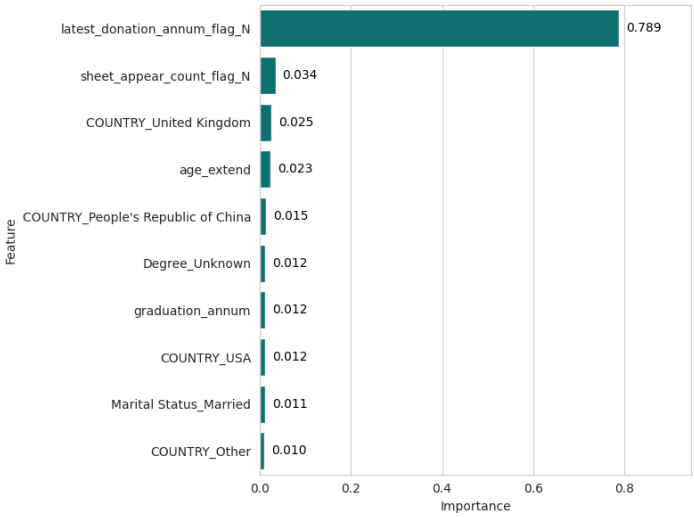


Figure 65: Feature Importance of Random Search XGBoost

6 Conclusion

This study identifies and compares the characteristics of potential legacy donors and pipeline participants in three approaches: unsupervised cluster analysis and factor analysis, as well as supervised predictive models. Due to the limitation of data provided, some of the results may not be satisfactory, but conclusions and implications can still be drawn to assist the PAGE Department with legacy donation analysis. Results of different methods may coincide or disagree, which are summarised and compared below in this section.

To start with, a comprehensive comparison is made based on the results of cluster analysis and prediction models. These methods generated lists of **potential donors** and **potential participants** based on model findings, which reached a general consensus. In pipeline dataset and active alumni dataset, the consistency rates for identifying positive samples (potential donors and participants) are 68.1% and 71% respectively. Two methods have obvious intersections in the result, but the difference could not be ignored either. In application, it is suggested to combine both methods and retain the positive samples of identification as much as possible so as not to miss any of the potential donors or participants. The PAGE department may send additional promotional emails to these alumni to retain their interests to donation or the legacy pipeline. Regarding the interpretation of the core features, both two methods provide valuable ideas for the legacy marketing strategy. The cluster analysis emphasizes group profile and behavior pattern recognition, which is suitable for formulating long-term cultivating strategies. The predictive model captures the level of positive engagement in donation participation, identifying active alumni and optimizing the resource input in the donation process. Both methods has their own advantages. The former provides potential population segmentation and characteristic profiling. The latter gives the prediction for the future successful conversion.

Besides, factor analysis and predictive models can both identify a list of variables which play significant roles in the datasets, though the former is concerned with latent structures in data, while the later concentrates on classification of two groups of alumni. Both methods successfully combat the curse of dimensionality and provide guidance to the PAGE Department with specific features that they might want to focus on to better understand potential donors and participants. Some significant variables identified by two methods align with each other. For instance, both methods suggest that `sheet_appear_count_flag` in **merged_pipeline**, as well as, `stage_enquirer` and `enquirer_annum` in **active_alumni_merge** are significant in describing potential donors or participants. However, predictive models identify age and previous donation records as important features, while factor analysis indicates enquiry-related features and degree are more influencing among collected data. Therefore, the PAGE department may require more information in such fields, under ethics consideration, to facilitate the analysis of legacy pipeline. In brief, no matter these results coincide or not, the insights brought could assist the PAGE Department with data collection and integration, which is likely to improve the quality of modeling and analysis explained in this study.

References

Ansari, A. and Jedidi, K. (2000), 'Bayesian factor analysis for multilevel binary observations', *Psychometrika* **65**(4), 475–496.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984), *Classification and Regression Trees*, Wadsworth.

Chen, T. and Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 785–794.

Cudeck, R. (2000), Factor analysis methods, *in* H. E. Tinsley and S. D. Brown, eds, 'RESEARCH in ORGANIZATIONS Foundations and Methods of Inquiry', Academic Press, pp. 265–296.

Friedman, J. H. (2001), 'Greedy function approximation: A gradient boosting machine', *Annals of Statistics* **29**(5), 1189–1232.

Green, C. L. and Webb, D. J. (1997), 'Factors influencing monetary donations to charitable organizations', *Journal of Nonprofit Public Sector Marketing* **5**(3), 19–40.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edn, Springer.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning: With Applications in R*, Springer.

Kolhede, E. and Gomez-Arias, J. T. (2022), 'Segmentation of individual donors to charitable organizations', *International Review on Public and Nonprofit Marketing* **19**, 333–365.

L., D.-C. P. and W., Z. (2012), 'A clusterwise linear regression model of alumni giving', *International Journal of Education Economics and Development* **3**(4), 330–347.

Le Blanc, L. A. and Rucks, C. T. (2009), 'Data mining of university philanthropic giving: Cluster-discriminant analysis and pareto effects', *International Journal of Educational Advancement* **9**(1), 64–82.

Lin, Y.-D. and Chang, Y.-L. (2009), 'Data mining of university philanthropic giving: Cluster-discriminant analysis and pareto effects', *International Journal of Educational Advancement* **9**(1), 64–82.

Pedro, I. M., Mendes, J. C. and Pereira, L. N. (2020), 'Identifying patterns of alumni commitment in key strategic relationship programmes', *Journal of Marketing for Higher Education* **30**(2), 1–20.

Quinn, K. M. (2004), 'Bayesian factor analysis for mixed ordinal and continuous responses', *Political Analysis* **12**, 338–353.

Rattanamethawong, R., Kanjanawattana, S. and Jermsittiparsert, K. (2016), 'An innovation model of alumni relationship management', *International Journal of Innovation, Creativity and Change* **2**(3), 1–15.

Sarmento, R. P. and Costa, V. (2019), 'Confirmatory factor analysis – a case study'. **URL:** *https://arxiv.org/abs/1905.05598*

Stewart, D. W. (1981), 'The application and misapplication of factor analysis in marketing research', *Journal of Marketing Research* **18**(1), 51–62.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.

Yang, B. (2005), Exploratory factor analysis, *in* R. A. Swanson and E. F. H. III, eds, 'Handbook of Applied Multivariate Statistics and Mathematical Modeling', Berrett-Koehler Publishers, Inc., San Francisco, pp. 181–199.

Önen, E. (2019), 'A comparison of frequentist and bayesian approaches: The power to detect model misspecifications in confirmatory factor analytic models', *Universal Journal of Educational Research* **7**(2), 494–514.