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

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 the following research aspects. The characteristics in various dimensions, such as age, address, year of class, department and degree, of people who have entered the pipeline are analysed and compared to the features of the whole active alumni population. The variability of characteristics among people in different stages and that between people who have withdrawn or not the programme are also investigated. These questions are addressed by a range of statistical and machine learning techniques including predictive models, factor analysis, and cluster analysis. Any potential changes over years in participants of the legacy pipeline are considered when models are developed and adaptations are made when necessary. 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

To be added after obtaining the final results....

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: predictive modeling, factor analysis, and cluster analysis.

Firstly, supervised learning models - logistic regression, decision trees, and random forests - are applied to predict the likelihood of alumni progressing through different stages of the legacy donation pipeline. These models provide both stage-specific conversion predictions and insight into the relative importance of key features.

Secondly, factor analysis technique is 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.

Thirdly, 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.

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

Following this, a numerical study is presented. This includes exploratory statistics of the overall alumni base and the legacy pipeline, visualisations of variable distributions, predictive model evaluation results (e.g., accuracy, AUC), factor loadings, and clustering outputs. These results form the empirical basis for subsequent interpretation.

Finally, we interpret the findings and provide practical recommendations for PAGE. These are designed 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 Predictive Models

2.1.1 Logistic regression

Logistic regression is a fundamental classification technique widely used in applied fields such as social science, medicine, and economics. Unlike linear regression, which models a continuous outcome, logistic regression is tailored for binary outcomes by modeling the probability that a given input belongs to a particular class.

James et al. (2013) emphasize logistic regression's interpretability and simplicity, particularly in scenarios where the relationship between predictors and a binary response variable is of interest. They demonstrate how logistic regression uses the logit function to map predicted values to the $[0, 1]$ interval, which is ideal for probabilistic classification.

Building on these concepts, Hastie et al. (2009) offer a more mathematical treatment of logistic regression. They explore the model's formulation via maximum likelihood estimation and connect it with generalized linear models (GLMs). Their work also analyzes the decision boundary created by logistic regression and its relationship with linear discriminant analysis (LDA) under specific distributional assumptions.

Recent studies have extended logistic regression in several directions. For instance, Ng and Jordan (2002) compare logistic regression with naïve Bayes for text classification tasks, showing that while naïve Bayes often performs better with limited data, logistic regression tends to outperform as the data volume increases. Additionally, Zou and Hastie (2005) introduce the elastic net, which combines L_1 and L_2 penalties to improve variable selection in the presence of correlated predictors.

In modern machine learning pipelines, logistic regression often serves as a baseline model due to its robustness and interpretability.

2.1.2 Decision trees

Decision trees are among the most interpretable and widely used non-parametric supervised learning models for both classification and regression tasks. Their structure resembles human decision-making, and they recursively partition the input space into distinct, homogeneous regions.

James et al. (2013) introduce decision trees as a flexible modeling tool that requires minimal data preprocessing, such as normalization or dummy variable creation. They emphasize the intuitive appeal of decision trees and discuss how models like classification and regression trees (CART) split predictor space based on features that most effectively reduce impurity, typically using metrics such as the Gini index or cross-entropy.

Hastie et al. (2009) provide a more theoretical and comprehensive analysis of decision trees, delving into their statistical underpinnings. They examine the trade-off between tree depth and model complexity and explain why fully grown trees often overfit training data. To address this, they introduce cost-complexity pruning, which balances model fit with tree size to improve generalization. Furthermore, they position decision trees as base learners

in ensemble methods like bagging and random forests, providing theoretical justification for their variance-reduction benefits.

Beyond these foundational texts, Breiman et al. (1984) formalized the CART methodology, which became the foundation of modern decision tree algorithms. Later, Breiman (2001) proposed random forests, an ensemble of decision trees that significantly improves predictive performance and robustness by aggregating outputs from multiple decorrelated trees trained on bootstrapped samples with randomly selected features.

Recent advancements have led to gradient boosting decision trees (GBDT), a technique introduced by Friedman (2001), which builds trees sequentially, each one correcting errors of the previous.

Despite their strengths, decision trees are prone to high variance, and their performance often improves when used in ensembles. Nevertheless, their interpretability and clear decision logic continue to make them valuable, especially in applications where model transparency is essential.

Model comparison. While both logistic regression and decision trees are used for binary classification, they offer distinct advantages depending on the modeling goals. Logistic regression is grounded in statistical inference and assumes a linear relationship between predictors and the log-odds of the response. Its coefficients are interpretable as odds ratios, making it ideal for understanding the individual effect of each variable on the outcome. In contrast, decision trees are non-parametric and excel at modeling non-linear relationships and interactions without requiring explicit specification. They produce intuitive flowchart-like structures that mirror human-like decision-making processes but are more prone to overfitting. In this report, both methods are implemented to leverage their respective strengths: logistic regression for interpretability and decision trees for capturing complex patterns in alumni behavior.Ensemble methods such as random forests enhance the predictive performance of decision trees, but they tend to be less interpretable.

2.2 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 maximum 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.

Although binary data can be handled either assuming it to be continuous or using Bayesian methods, the algorithms are still computationally intensive, which could be improved by using limited information estimation instead of relying on the full joint distribution of all observed variables. Wu and Bentler (2013) reviews and extends the use of limited information estimators in factor analysis, including the weighted least squares and multinomial maximum likelihood methods with second and third order moments. In comparison to full-information estimators, the authors suggests that the third-order moments weighted least squares estimator performs the best in estimating factor loadings with the lowest overall root-mean-square error (RMSE) and generally good fit of the model, while the Laplace approximation maximum likelihood estimator generates the most accurate intercept estimation, which provides a viable way for complex models using full information. However, MCMC demonstrates slow convergence and the least accurate estimations in this research.

In addition to continuous and binary data, categorical variables are also commonly used in factor analysis, either ordinal or nominal. da Silva et al. (2020) proposes Bayesian factor analysis models for categorical data obtained from questionnaires using MCMC for parameter estimation, in which the ordinal data is considered as discretised versions of underlying continuous response and modeled by a cumulative logistic regression model, and the latent variable behind the nominal data is modeled by a multinomial logistic regression. The effectiveness of this method for nominal factor analysis is also emphasized by Revuelta et al. (2020) with various applications in psychology, education, and social science research, which is especially suitable for analysing the first choice data, where an option is chosen by respondents from a set of unordered alternatives.

2.3 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 un-

derstand 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, universities will be better positioned to distinguish alumni attitudes and behaviors, establish stronger relationships and increased philanthropy.

3 Methodology

3.1 Predictive Models

In this part, we focus on three widely used classification algorithms — logistic regression, decision trees, and random forests — to model alumni data in relation to legacy doantion. These models are selected due to their strong presence in the literature and their complementary strengths: logistic regression offers interpretability and statistical inference, decision trees capture non-linear relationships and feature interactions, and random forests enhance predictive accuracy through ensemble learning.

The prediction task is formulated as a binary classification problem, where the objective is to identify alumni who are likely to engage in legacy giving. While the legacy pipeline consists of multiple stages (e.g., Enquirer, Intender, Pledge, Proposal), model is not trained on each transition individually. Instead, focus is assigned on predicting overall engagement with PAGE, while complementing the analysis with descriptive insights across pipeline stages. This approach balances predictive power with interpretability and allows us to extract insightful and actionable findings relevant to different stages of donor development.

To implement this framework, three models are trained using a labeled dataset that includes demographic, academic, and behavioral information about LSE alumni. Key predictors include graduation year, degree type, field of study, prior donation history, engagement with outreach campaigns, and estimated income bracket. These features are preprocessed and standardized as necessary for each model. The following sections describe the implementation and rationale for each model.

3.1.1 Logistic Regression

The logistic regression model is estimated via maximum likelihood, modeling the log-odds of a positive outcome as a linear function of the predictors:

$$\log\left(\frac{P(Y=1 \mid \mathbf{X})}{1-P(Y=1 \mid \mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{1}$$

Categorical variables are one-hot encoded, and continuous variables are standardized prior to estimation. To account for potential multicollinearity and overfitting, regularization techniques such as L1 (lasso) or L2 (ridge) penalties may be incorporated. For example, L2-regularization modifies the objective function as follows:

$$\mathcal{L}(\boldsymbol{\beta}) = -\log L(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2}$$

The regularization strength λ is tuned using cross-validation.

3.1.2 Decision Trees

Decision trees are implemented using the CART algorithm. A decision tree recursively partitions the predictor space into subsets that are increasingly homogeneous with respect to

the outcome variable. At each node, the algorithm selects the feature and threshold that minimize a predefined impurity measure.

For classification tasks, Gini impurity is commonly used as the splitting criterion. In the case of binary classification, where the two classes are denoted as class 0 and class 1, the Gini index at a given node t is calculated as:

$$G(t) = 2p(1-p) \tag{3}$$

where p is the proportion of class 1 samples at node t. The Gini impurity reaches its minimum (zero) when all samples in the node belong to a single class and is maximized (at 0.5) when classes are evenly split. The tree continues to grow until a stopping criterion is met, such as a minimum number of samples per leaf or maximum tree depth.

3.1.3 Random Forest

The random forest model builds an ensemble of decision trees trained on different bootstrap samples of the data. At each split, only a randomly selected subset of features is considered, which introduces diversity among the trees and reduces overall variance (Breiman, 2001).

For prediction, the final output is determined by majority vote across all trees:

$$\hat{y} = \text{majority\_vote}\left(h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_T(\mathbf{x})\right) \tag{4}$$

Alternatively, for probabilistic predictions, the average predicted probability from all trees can be used:

$$\hat{P}(Y = 1 \mid \mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} h_t(\mathbf{x}) \tag{5}$$

The number of trees, maximum depth, and number of features considered at each split are tuned using grid search with cross-validation. Feature importance is derived from the average decrease in Gini index across trees, providing insights into which variables are most influential at different stages of the donation process.

Model Evaluation

Model performance is evaluated using metrics including accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (AUC-ROC). Accuracy measures the overall proportion of correct predictions, while precision and recall provide insights into the model's ability in predicting the positive class, particularly in imbalanced settings. The F1 score represents the harmonic mean of precision and recall, balancing both false positives and false negatives. AUC-ROC summarizes the model's ability to discriminate between classes across different classification thresholds. These metrics are used to compare the effectiveness and trade-offs of different models.

3.2 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 given as below.

$$X_{p,i} - \mu_p = \lambda_{p,1}F_{1,i} + \lambda_{p,2}F_{2,i} + \cdots + \lambda_{p,k}F_{K,i} + \epsilon_{p,i}$$

where $\lambda p, k$ is the loading for the variable X_p on factor F_k,

μ_p is the mean of X_p, and

$\epsilon_{p,i}$ has zero mean and finite variance.

Equivalently, in matrix notation

$$\mathbf{X} - \mathbf{M} = \mathbf{LF} + \epsilon$$

where $\mathbf{X} \in \mathbb{R}^{P \times n}$ is the observation matrix,

$\mathbf{M} \in \mathbb{R}^{P \times n}$ is the matrix of variable means,

$\mathbf{L} \in \mathbb{R}^{P \times K}$ is the factor loading matrix,

$\mathbf{F} \in \mathbb{R}^{K \times n}$ is the factor matrix, and

$\epsilon \in \mathbb{R}^{P \times n}$ is the error matrix

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.

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top + \mathbf{\Psi}$$

where $\mathbf{L}\mathbf{L}^\top$ quantifies the shared variances, and

$\mathbf{\Psi} \in \mathbb{R}^{P \times P}$ is the diagonal matrix of unique variances

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.

In EFA, coefficients of hidden factors are first estimated via extraction methods such as principal components analysis, principal axis factoring and canonical factor analysis. The number of factors should be used can be selected by different methods including Horn's parallel analysis, MAP test, and Kaiser criterion. 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. While oblique rotation might be more suitable when factors are correlated, and so is more realistic in real-world applications.

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 then can be assessed by chi-square tests, root mean square error of approximation, comparative fit index.

3.3 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.

This project begins by assembling a cleaned, preprocessed dataset comprising alumni engagement activity (event attendance, volunteering, email interaction, donation history), academic background (programme group, department, graduation year), and inferred demographic and behavioural indicators (e.g., inferred income via country GDP at graduation). Features are constructed based on theoretical relevance and data availability, with particular attention to four hypothesised drivers of legacy giving: personal wealth, institutional affinity, student experience, and peer influence. Categorical variables such as programme group and region are one-hot encoded, while skewed continuous variables such as number of

events attended or email clicks are log-transformed to reduce the influence of extreme values. All features are standardised to zero mean and unit variance to ensure fair contribution to distance calculations. Highly correlated features (correlation ¿ 0.8) are pruned to mitigate multicollinearity.

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, and Bayesian Gaussian mixture models (BGMM). 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. This study will experiment with a range of k values (e.g., 3 to 10), selecting the optimal number of clusters using validation metrics such as silhouette score, Calinski–Harabasz index, and the elbow method. K-means' speed makes it suitable for running on the full dataset or on stratified samples, enabling us to capture broad engagement patterns and facilitate communication with non-technical stakeholders.

However, alumni engagement patterns may not conform to simple spherical clusters. Therefore, hierarchical agglomerative clustering will also be implemented, which builds a nested tree of clusters by iteratively merging the most similar pairs based on a linkage criterion (such as Ward's method). 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. Hierarchical clustering is particularly effective for identifying nested substructures within the data—e.g., differentiating between alumni who attend events occasionally versus those who attend regularly but do not donate. Although it is computationally more expensive, it provides valuable visual insights through dendrograms and is useful for exploratory analysis and validation of k-means results.

A third technique proposed is Bayesian Gaussian mixture modelling. Unlike k-means, which assigns each data point to a single cluster, BGMM assumes that each cluster is a Gaussian distribution and that observations may belong to multiple clusters with varying probabilities. This is especially important in our context, where alumni behaviours may not be rigidly separable. For instance, an individual may attend events regularly but only donate after retirement, displaying traits of both engaged participants and late-stage donors. 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 gives soft assignments—probabilistic representations of clusters—to each alum, and this 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.

To compare and evaluate clustering solutions, internal as well as external measures will be used. Internal measures like cluster compactness will be calculated for raw and boot-

strapped samples to verify robustness and stability of the cluster solution. The clusters will be visualized through dimension reduction techniques like t-SNE or UMAP to examine cluster separability in lower dimension. Following the clustering process, every cluster will be described with average feature values, demographic profile, and behavior, determining significant segments like "Affluent but Disengaged," "Young High-Potential Alumni," or "Peer-Driven Regional Donors."

Apart from profiling, the relationship between the clusters and pipeline progression in the past will be established through their inclusion in predictive models. That is, cluster affiliation (hard or soft memberships) as a predictor will be incorporated in logistic regression or survival analysis for the prediction of the probability of progression from Passive to Transactional, or from Participant to Partner. Temporal stability for the clusters will be established through the use of metrics like the Population Stability Index (PSI) and the Adjusted Rand Index (ARI) in comparing cluster assignments across fiscal years. If the clusters change considerably, retraining intervals will be recommended (e.g., every 3–5 years) in preparation for achieving predictive power.

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 but unengaged clusters, for example, may be highlighted as being deserving of proactive solicitation, and low-return segments deprioritized. Further, the BGMM's soft cluster assignments can be used to probabilistically rank alumni for solicitation for the purpose of the legacy campaign to direct the limited available fundraising resources more effectively.

Simply put, clustering is a powerful method of extracting meaning from high-dimensional, complex alumni activity data. By using multiple cluster methods—each with its own particular strength—what consistently find is the underlying patterns of behavior that inform pipeline development. 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.

4 Modeling

5 Conclusion

References

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

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

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.

da Silva, V. G. C., Gonçalves, K. C. M. and Pereira, J. B. M. (2020), 'Bayesian factor models for multivariate categorical data obtained from questionnaires'.
URL: *https://arxiv.org/abs/1910.04283*

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*, 2 edn, Springer, New York.

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

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.

Ng, A. Y. and Jordan, M. I. (2002), On discriminative vs. generative classifiers: A comparison of logistic regression and naïve bayes, *in* 'Advances in Neural Information Processing Systems (NeurIPS)', pp. 841–848.

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.

Revuelta, J., Maydeu-Olivares, A. and Ximénez, C. (2020), 'Factor analysis for nominal (first choice) data', *Structural Equation Modeling: A Multidisciplinary Journal* **27**(5), 781–797.

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.

Wu, J. and Bentler, P. M. (2013), 'Limited information estimation in binary factor analysis: A review and extension', *Computational Statistics Data Analysis* **57**(1), 392–403.

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.

Zou, H. and Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.

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.