Survivability of charitable organisations – What factors influence whether a charity survives?

(This dissertation was undertaken as part of the CDRC Masters Dissertation Scheme)

Vaishnavi Patil
201896641

Supervised by :
Dr. Luissa Cutillo (University of Leeds)
Christopher Davy (Social Investment Business)
Alannah Keogh (Social Investment Business)

Submitted in accordance with the requirements for the
module MATH5872M: Dissertation in Data Science and Analytics
as part of the degree of

Master of Science in Data Science and Analytics

The University of Leeds, School of Mathematics

August 2025

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

UNIVERSITY OF LEEDS

School of Mathematics

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Academic integrity statement

I am aware that the University defines plagiarism as presenting someone else's work, in whole or in part, as your own. Work means any intellectual output, and typically includes text, data, images, sound or performance.

I promise that in the attached submission I have not presented anyone else's work, in whole or in part, as my own and I have not colluded with others in the preparation of this work. Where I have taken advantage of the work of others, I have given full acknowledgement. I have not resubmitted my own work or part thereof without specific written permission to do so from the University staff concerned when any of this work has been or is being submitted for marks or credits even if in a different module or for a different qualification or completed prior to entry to the University. I have read and understood the University's published rules on plagiarism and any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all my work in any form and using third parties (who may be based outside the EU/EEA) to monitor breaches of regulations, to verify whether my work contains plagiarised material, and for quality assurance purposes. I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the University's policy on mitigation and the School's procedures for the submission of statements and evidence of mitigation. I am aware of the penalties imposed for the late submission of coursework.

Name : Vaishnavi Patil

Student ID : 201896641

Abstract

The UK charity sector comprises high number of registered organisations that collectively deliver substantial social and economic value, yet face growing sustainability challenges. Despite this importance, the determinants of organisational survival remain insufficiently understood, with past research relying on small samples or U.S centred frameworks.

This dissertation addresses the question: *What factors determine charitable organisation survival in the UK, and how can predictive modelling support evidence-based decision-making in the sector?* Using full administrative records from the Charity Commission of England and Wales (1960–2024, 8.03 million records across 13 tables), the study presents the large-scale empirical analysis of UK charity survival.

A structured methodological pipeline transformed administrative data into features spanning governance, finance, structure, temporal compliance, mission clarity, and environment. Logistic regression was adopted for its interpretability in policy contexts. The model achieved 81.6% accuracy and 99.0% recall in identifying at-risk organisations.

Results identified five critical predictors of survival: governance quality (the strongest determinant), younger organisational age, meeting financial thresholds above £10,000, operational recency, and modern legal forms such as Charitable Incorporated Organisations. Findings challenge aspects of traditional lifecycle theory by revealing greater short-term resilience among younger charities. The study provides both theoretical contribution and practical tools, including a prototype web application for real-time survival assessment, enabling funders, regulators, and managers to strengthen sector sustainability.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Context

The charity sector is a massive part of England's civil society, making a contribution worth billions of pounds to the economy as they address key aspects of social needs. It embraces organizations running from small hardly resourced community groups up to large national charities possessing massive resources and complex governance structures. Though their function is hinged on delivering an important social mission, most charitable organizations are confronted with great sustainability problems that endanger the essence of their operations and service delivery toward beneficiaries.

The reason why certain charities continue to exist while others go out of business has risen to significant relevance amid growing financial pressures on the sector, changes in regulatory requirements, and evolving donor expectations. The COVID-19 pandemic has served to underscore even more the precarious position many charitable organizations find themselves in with some suffering drastic cuts in funding while being called upon to do even more work. Such a difficult environment makes it not only interesting but imperative that one tries to determine those factors that would go into organizational survival and then develop tools that go into predicting which charities are likely to be closed down.

The UK charity sector works under an established regulatory environment supervised by the Charity Commission for England and Wales. From this regulatory setting, there is developed detailed administrative data on charities that cover aspects related to their finances, governance, operations, and legal standing. This rich dataset opens up a hitherto unavailable chance of probing into the patterns of survival of charities through advanced analytical techniques complemented by large-scale empirical analysis.

Recent advances in data science and machine learning enable large-scale testing of organizational survival theories in the nonprofit sector. Unlike previous studies using small samples, comprehensive administrative data and advanced analytics now allow rigorous exploration of charity survival factors. This creates opportunities to develop predictive models supporting evidence-based decision-making for charity managers, funders, and policymakers.

1.2 Problem Statement and Research Gap

Charities are very important to the community and, despite their struggles to keep afloat, receive little attention in research that would use large-scale empirical data to analyse the issue of survival, particularly in the UK. Most existing nonprofit survival studies have used data from the US, where regulatory environments, funding mechanisms, and other organizational attributes differ so markedly from those found within the UK charity sector.

A handful of studies that consider UK charities mostly discuss some particular element of organizational performance or work with comparatively small datasets through which it is difficult to generalize findings. There is an evident lack of research that brings together a broad exploratory analysis of charity characteristics and advanced forms of predictive modelling to explore patterns in survival.This limits the ability of sector stakeholders to make informed decisions about organizational development, funding allocation, and policy interventions.

Earlier research often focused narrowly on financial indicators as the primary measure of charity survival. However, this approach can overlook important organisational, governance, and operational factors that may be equally or even more influential in determining long-term sustainability. This study responds to that gap by highlighting the need for broader methodologies that analyse how multiple organisational characteristics interact to shape survival outcomes.

The lack of robust predictive tools for charity survival also represents a practical problem for sector stakeholders. Charity managers need better understanding of the factors that contribute to organizational sustainability to guide strategic planning and operational decisions. Funders require tools to assess the long-term viability of potential grant recipients. Policymakers need evidence about which regulatory and support interventions are most effective in promoting sector sustainability.

1.3 Research Aim and Questions

The leading focuses of this study are to increase perceptions of charitable group survivability through complete information science evaluation, and to create evidence-based apparatuses that might fortify the manageability of the UK charity area. This examination is the primary huge scope experimental request of charity endurance involving the whole UK regulatory dataset. It adds both to academic comprehension of nonprofit authoritative elements and to down-to-earth area leaders' capacity.

This study addresses the fundamental question: What factors determine charitable organisation survival in the UK, and how can predictive modelling support evidence-based decision-making in the charity sector?

To explore this comprehensively, the research examines five specific dimensions of charity survival through a balanced approach combining exploratory data analysis with advanced technical methodology:

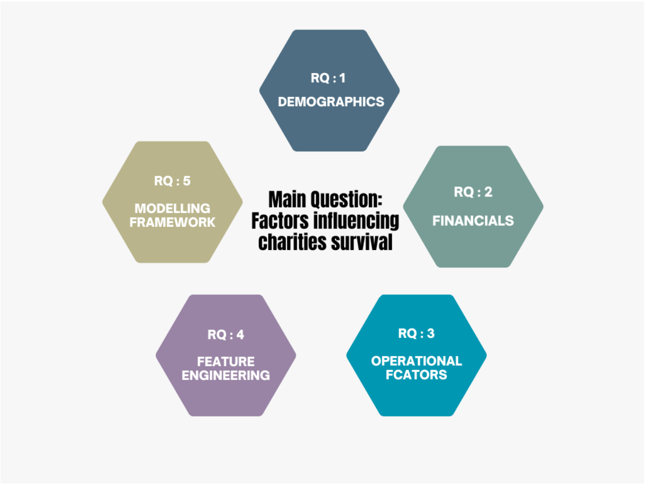


Figure 1.1: Research Framework: Five-Question Analytical Structure

RQ1: What organizational demographics and structural characteristics differentiate surviving charities from those that cease operations? This question focuses on identifying key demographic patterns, legal structures, geographic scope, and registration trends that distinguish successful organizations from those that cease operations. The analysis examines over 390,000 charitable organizations to uncover systematic differences in organizational foundations and structural characteristics.

RQ2: How do financial characteristics and resource patterns influence survival probability across charity income segments? This investigation explores the relationship between financial capacity, income thresholds, funding diversification, and organizational sustainability. The research examines how resource availability and financial management strategies impact survival outcomes across different charity sizes and sectors.

RQ3: Which operational factors exhibit the strongest association with charity survival outcomes? This analysis focuses on understanding the most common charity classifications.

RQ4: How can advanced feature engineering extract predictive signals from charity administrative data to improve survival prediction accuracy? This addresses the technical challenge of transforming raw administrative data into meaningful predictive features. The research employs sophisticated data science techniques to uncover hidden patterns and create composite indicators that enhance model performance while revealing non-obvious survival predictors.

RQ5: What modelling approach and validation framework optimizes both predictive performance and interpretability for charity sector decision-making applications? This investigation focuses on developing robust predictive models that balance statistical rigor with practical interpretability. The research emphasizes comprehensive validation frameworks and model selection approaches that ensure reliable deployment in real-world charity sector applications.

The research is designed to bridge the gap between academic analysis and practical application, ensuring that sophisticated data science techniques generate actionable insights for strengthening charity sector sustainability.By combining rigorous analytical methods with comprehensive empirical data, study aims to provide both theoretical contributions to nonprofit research and practical tools that can support strategic decision-making across the charitable sector.

1.4 Scope and Significance

The study of registered charities in England and Wales, using full administrative data from the Charity Commission for the whole period from registration to 2024. More than 390,000 charitable organizations are included in this analysis.This scope already provides sufficient statistical power to identify robust patterns and develop reliable predictions. This enables an assessment of how survival patterns may have changed with transformations in regulatory frameworks, funding environments, and even operational practices. Such a timeframe not only helps in understanding trends from a historical perspective but also brings out current patterns that might be useful for future interventions in the sector.

The study uses binary survival outcomes, labelling organizations as either ongoing concerns or those that have exited operations (regulatory status). Though such a simplistic reduction of the complex reality of change and development in organizations to a binary outcome may be inappropriate, it does provide an unambiguous framework for predictive modelling. Also, this provides direct practical use.

This study is therefore significant to numerous stakeholder groups inside and outside of the charity sector. It advances scholarly understanding of nonprofit organizational survival and, in practical terms, shows how advanced forms of analysis can be fruitfully applied to existing administrative data. In addition, while most work in this area reports on findings from survival modelling within the US context, these results are based on modelling within the UK context.

For charity practitioners and managers, this research comes up with insights based on evidence regarding organizational characteristics and operational strategies that contribute to sustainability. This predictive modelling framework can be used in strategic planning, risk assessment, and organizational development decisions. The outcomes of the study identified critical factors such as governance quality and financial thresholds that will now inform and direct resource allocation and capacity building efforts.

It arms funders and investors in the charity sector with tools for assessing the long-term viability of potential grant recipients and understanding which types of organizational support are most likely to promote sustainability. The ability to predict survival outcomes would no doubt inform funding strategies and thus implementation of resource allocation, particularly when resources are very scarce.

This research provides policymakers with insights on effective regulatory interventions to promote sector sustainability. It helps inform policy, regulatory reform, and sector support programs, while identifying emerging risks and opportunities. The focus on model interpretability and the use of administrative data ensures the findings are actionable and can be easily updated with new data.

Chapter 2

Literature Review

2.1 Introduction to the Literature

Understanding charity survival factors has gained increasing significance amid financial stringency and regulatory reforms in the UK charity sector. Literature on charity survival draws from organizational ecology, financial management, and nonprofit governance studies, assessing factors influencing organizational resilience and continuation during crises with specific reference to UK contexts and methodological approaches.

Early works focused on financial indicators of organizational distress using commercial sector bankruptcy models. However, charitable organizations unique characteristics mission oriented operations, diversified funding sources, and distinct regulatory environments prompted researchers to develop specialized approaches. This evolved literature now incorporates organizational characteristics, governance quality, operational strategies, and environmental factors alongside traditional financial measures. The UK charity sector, with over 170,000 registered bodies, provides an ideal setting for examining survival patterns given comprehensive administrative data availability and significant economic and social impact.

2.2 Theoretical Framework: Financial Theory

The foundational framework for nonprofit financial vulnerability was established by Tuckman & Chang (11), who identified four critical financial indicators that predict organizational distress: revenue concentration, inadequate operating reserves, unusually low administrative costs, and declining revenue patterns. Their groundbreaking study demonstrated that organizations exhibiting multiple vulnerability indicators faced significantly higher dissolution risk. Unlike commercial bankruptcy models, the Tuckman-Chang framework recognized that charitable organizations operate under fundamentally different financial constraints, including donor restrictions and mission obligations that shape financial management strategies.

Building upon this foundation, Green et al. (2) advanced the field through their analysis of financial resilience in UK charities, introducing four critical dimensions: revenue stability patterns, organizational reserve adequacy, operational cost efficiency, and adaptive capacity during financial stress. Their empirical analysis demonstrated that organizations with higher financial resilience scores experienced significantly lower dissolution rates. The study particularly emphasized income diversification strategies and reserve management practices as key protective factors against organizational failure in the UK charity context.

Searing et al. (7) further refined the framework by developing the nonprofit resiliency framework, examining how organizations actively build financial resilience during crisis periods. Their approach identified three core dimensions: financial resource management tactics that optimize cash flow during stress periods, revenue diversification strategies that reduce single source funding dependence, and crisis response mechanisms enabling rapid adaptation to changing circumstances. This framework emphasizes organizational agency in building resilience rather than simply identifying vulnerability factors.

The integration of these frameworks creates a comprehensive approach to charity financial sustainability, progressing from vulnerability identification through resilience building to proactive crisis management. This theoretical evolution provides the foundation for empirical analysis of UK charity survival factors.

2.3 Organizational Factors: Age, Structure, and Size Effects

Organizational characteristics represent fundamental predictors of charity survival, with age, structure, and size forming key determinants of organizational longevity. Green et al. (2) provide contemporary evidence on how these organizational factors interact with financial resilience to influence survival outcomes in UK charities. Their systematic analysis demonstrates that organizational age effects must be understood alongside structural characteristics and size-related financial capacity.

The relationship between organizational maturity and survival has evolved beyond traditional liability of newness predictions. Green et al. (2) show that while newer organizations face establishment challenges, age advantages are not automatic and depend significantly on how organizations develop their income structures over time. Established organizations may face survival risks if they become overly dependent on traditional income sources without adapting to changing funding environments.

Size remains a critical survival determinant, with **?**) demonstrating that financial capacity significantly influences nonprofit sustainability. Larger organizations typically maintain greater resource buffers, more diversified revenue streams, and enhanced administrative capacity—factors that improve survival prospects. However, Green et al. (2)'s findings suggest that size effects are mediated by income composition, with smaller organizations achieving sustainability through strategic resource management rather than scale alone.

Structural characteristics, particularly legal form, have become increasingly significant with new organizational forms in the UK charity sector. Charitable Incorporated Organizations

(CIOs) combine incorporation advantages with simplified regulatory requirements compared to traditional charitable trusts and charitable companies. Understanding how different legal structures influence survival outcomes has practical implications for charity formation decisions and regulatory policy development.

2.4 Governance Quality and Operational Characteristics

Good governance has increasingly emerged as crucial for charity survival as researchers move beyond purely financial approaches to organizational sustainability. Herman & Renz (3) identify governance quality as a determinant factor for long-term capacity, encompassing multidimensional aspects including board composition, strategic oversight, financial management, and regulatory compliance.

For the purposes of empirical measurement, governance quality is typically operationalized through structural indicators such as board size and composition, with adequate governance generally defined as having between 5-10 trustees and strong governance as having more than 10 trustees, reflecting the capacity for effective oversight and decision-making.

Strong governance manifests through effective financial management, strategic decision-making, and high stakeholder confidence. Governance quality provides superior survival mechanisms compared to immediate financial indicators because it creates conditions for sustainable financial management over time. Assessment should consider both structural elements (board size and composition) and process elements (strategic planning and performance monitoring).

Hyndman & McDonnell (5) provide a comprehensive framework for understanding charity governance as relating to ”the distribution of rights and responsibilities among and within the various stakeholder groups involved, including the way in which they are accountable to one another; and also relating to the performance of the organization, in terms of setting objectives or goals and the means of attaining them.” Their stakeholder-based approach demonstrates how governance quality encompasses both internal mechanisms (board composition, staff relations) and external relationships (donor accountability, regulatory compliance).

Revenue diversification strategies represent another crucial operational characteristic affecting survival. Lu et al. (6) provided comprehensive empirical validation through meta-analysis of 258 effect sizes from 23 studies. They confirmed that revenue diversification increases financial capacity and reduces vulnerability across organizational contexts, while identifying important moderating factors including organizational size, sector focus, and environmental conditions that influence diversification strategy effectiveness.

2.5 Geographic and Socioeconomic Influences

Geographic context influences charity survival through multiple mechanisms including local economic conditions, demographic characteristics, and competition for resources. The UK presents a particularly interesting context for examining geographic influences given the distinct regulatory and cultural environments between England and Wales, as well as significant regional variations in economic conditions and charity sector development.

Green et al. (2) provided crucial insights into UK charity survival patterns through their analysis of financial resilience and organizational survival in UK charities. Their study represents one of the few comprehensive examinations of survival factors using UK administrative data. They found that income dependence patterns and financial resilience measures were significant predictors of organizational survival, with important variations across different types of organizations and geographic areas.

Local economic conditions directly affect charitable organizations through their impact on individual and corporate giving capacity, demand for charitable services, and availability of volunteers. Organizations operating in economically disadvantaged areas may face particular sustainability challenges due to limited local funding opportunities combined with high service demand. Conversely, charities in affluent areas may benefit from enhanced funding availability but face different competitive pressures and higher operational costs.

The scope of geographic operations whether organizations focus on local communities or operate nationally also influences survival prospects. Local organizations may benefit from stronger community connections and lower competition for funding, while national organizations may achieve greater revenue diversification and economies of scale. Understanding these tradeoffs is important for strategic planning and organizational development decisions.

2.6 Predictive Modelling in Charity Survival Research

Predictive modelling in charity survival research has evolved from basic financial ratio analysis to sophisticated statistical approaches that balance accuracy with interpretability. Contemporary applications demonstrate the practical value of well established statistical methods for addressing real-world nonprofit management challenges.

Song (10) demonstrates the application of binary logistic regression in nonprofit research through analysis of 8,408 American charities, using this methodology to predict organizational accountability and transparency scores from financial and operational indicators. While focused on accountability rather than survival, this large-scale empirical study establishes logistic regression's continued relevance for nonprofit prediction models. Recent methodological advances by Searing (8) have extended traditional nonprofit models to hybrid organizations, demonstrating how established frameworks can accommodate emerging organizational forms while maintaining interpretability essential for practical application.

The choice between logistic regression and more complex machine learning techniques involves important trade-offs for charity survival research. While advanced methods may achieve marginally higher predictive accuracy, logistic regression maintains significant advantages for research focused on understanding causal relationships and informing policy decisions. The transparency of logistic regression coefficients provides crucial benefits for both academic understanding and practical application, allowing researchers and practitioners to identify which specific factors drive survival outcomes and by what magnitude.

Hosmer et al. (4) provide comprehensive theoretical foundations for logistic regression in binary outcome research, establishing the mathematical basis for coefficient interpretation, model diagnostics, and statistical inference that makes this approach particularly suitable for organizational survival analysis where stakeholders require interpretable results for decision-making.

2.7 Methodological Foundations

The empirical analysis of charitable organization survival requires sophisticated statistical methods capable of handling complex administrative data while providing interpretable results for policy applications. This section establishes the mathematical and theoretical foundations for the analytical approaches employed in charity survival research.

2.7.1 Survival Analysis for Organizational Research

Survival analysis is a statistical methodology that examines the time until a specific event of interest occurs, typically referred to as "failure" or the "event." In organizational contexts, survival analysis investigates the duration until organizations cease operations, with the survival function $S(t)$ representing the probability that an organization survives beyond time t:

$$S(t) = P(T > t) \tag{2.1}$$

where T represents the random variable denoting time to organizational failure.

Binary Survival Analysis Framework

Binary survival analysis simplifies the temporal complexity by examining organizational status at a fixed observation point, creating dichotomous outcomes where organizations are classified as either surviving (1) or having failed (0). This approach transforms the continuous time-to-event problem into a binary classification framework:

$$Y_i = \begin{cases} 1 & \text{if organization } i \text{ survives to observation point} \\ 0 & \text{if organization } i \text{ fails before observation point} \end{cases} \tag{2.2}$$

2.7.2 Logistic Regression Framework

Logistic regression has emerged as the standard analytical approach for binary organizational survival outcomes due to its mathematical properties and interpretability advantages. The method provides both robust statistical inference and practical coefficient interpretation suitable for organizational stakeholders.

For organization i with predictor variables $\mathbf{X}_i$, the logistic regression model specifies the probability of survival as:

$$P(Y_i = 1|\mathbf{X}_i) = \frac{\exp(\boldsymbol{\beta}^T \mathbf{X}_i)}{1 + \exp(\boldsymbol{\beta}^T \mathbf{X}_i)} \tag{2.3}$$

The logit transformation enables linear modeling of the relationship between predictors and the log-odds of survival:

$$\text{logit}(P(Y_i = 1)) = \ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \boldsymbol{\beta}^T \mathbf{X}_i \tag{2.4}$$

This transformation ensures predicted probabilities remain within valid bounds while enabling interpretable linear relationships. Coefficient exponentials provide odds ratios: $\text{OR}_j = \exp(\beta_j)$, representing multiplicative changes in survival odds for unit increases in predictor X_j. For example, an odds ratio of 2.5 indicates that organizations with that characteristic are 2.5 times more likely to survive, providing intuitive interpretation for practitioners and policymakers (4).

2.7.3 Feature Selection Methods

Effective feature selection represents a critical component in organizational survival research using high-dimensional administrative data. Multiple complementary methods enable identification of informative organizational characteristics while maintaining statistical validity.

Correlation Analysis

Correlation measures the strength and direction of linear association between two variables. The Pearson correlation coefficient r between variables X and Y is defined as:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} \tag{2.5}$$

where $\bar{X}$ and $\bar{Y}$ represent sample means. The correlation coefficient ranges from -1 to $+1$, where:

- $r = +1$ indicates a perfect positive linear relationship,
- $r = -1$ indicates a perfect negative linear relationship,
- $r = 0$ indicates no linear relationship.

Mutual Information

Mutual Information quantifies the information shared between predictor variables and survival outcomes, capturing both linear and non-linear relationships:

$$I(X;Y) = \sum_{x} \sum_{y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{2.6}$$

This approach proves particularly valuable for administrative data where relationships may be non-linear or threshold-based.

Statistical Hypothesis Testing Framework

In this study, statistical hypothesis testing is used to test relationships between organisational features and survival. The framework encompasses several key components:

Null Hypothesis (**H_0**): States no relationship exists between predictor and outcome variables.

Alternative Hypothesis (**H_1**): Proposes a significant relationship exists.

Test Statistic: For categorical predictors, the ANOVA F-statistic compares group variances:

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} = \frac{\frac{\sum n_i(\bar{y}_i - \bar{y})^2}{k-1}}{\frac{\sum\sum(y_{ij} - \bar{y}_i)^2}{N-k}} \tag{2.7}$$

where k is the number of groups, N is total sample size, and MS denotes mean squares.

p-value: The probability of observing results as extreme as those obtained, assuming H_0 is true.

Type I Error (**α**): False positive rate, typically set at 0.05, representing the probability of rejecting a true null hypothesis.

2.7.4 Alternative Modelling Approaches

While logistic regression serves as the primary analytical framework, alternative approaches provide methodological validation and assessment of interpretability-accuracy trade-offs.

Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees through bootstrap aggregation (bagging). The algorithm constructs B bootstrap samples from the original dataset and trains a decision tree on each sample, using only a random subset of m features at each split (typically $m = \sqrt{p}$, where p is the total number of features).

The final prediction combines individual tree predictions:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{2.8}$$

where $T_b(x)$ represents the prediction from the b-th tree for input x.

Random Forest demonstrates particular advantages for organizational data analysis: it naturally handles mixed data types (categorical and continuous variables common in administrative data), captures non-linear relationships without explicit specification, provides built-in variable importance measures, and maintains robustness to outliers through bootstrap aggregation.

However, Random Forest sacrifices interpretability for predictive performance, making coefficient-level insights impossible—a critical limitation for policy-oriented nonprofit research requiring transparent decision-making rationale.

Regularized Regression Techniques

Regularization addresses overfitting by adding penalty terms to the standard regression objective function. For logistic regression, the regularized objective becomes:

$$L(\beta) = -\sum_{i=1}^{n} \big[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\big] + \lambda\, \Omega(\beta) \tag{2.9}$$

where $\Omega(\beta)$ is the regularization penalty and λ controls regularization strength.

L1 Regularization (Lasso): Uses penalty $\Omega(\beta) = \sum_j |\beta_j|$, driving some coefficients to exactly zero, effectively performing automatic feature selection while maintaining interpretability.

L2 Regularization (Ridge): Uses penalty $\Omega(\beta) = \sum_j \beta_j^2$, shrinking coefficients toward zero without elimination, stabilizing solutions when predictors are highly correlated.

Both techniques prove valuable when dealing with high-dimensional administrative datasets where feature correlation and overfitting risks are substantial concerns for model generalization.

2.7.5 Model Validation Framework

Robust validation frameworks ensure model reliability and generalizability, particularly critical for policy applications requiring dependable predictions about organizational sustainability.

K-fold Cross-Validation Theory

K-fold cross-validation provides the theoretical foundation for assessing model generalization while preventing overfitting. The method partitions the dataset into k subsets, training on $k - 1$ folds and testing on the remaining fold, repeated k times:

$$CV(k) = \frac{1}{k} \sum_{i=1}^{k} L(y_i, \hat{f}^{(-i)}(x_i)) \tag{2.10}$$

where $\hat{f}^{(-i)}$ represents the model trained excluding fold i, and L is the loss function.

Performance Assessment Metrics

Performance assessment employs multiple complementary measures derived from confusion matrix elements:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.11}$$

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{2.12}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.13}$$

Area Under the ROC Curve (AUC) provides threshold-independent discrimination assessment:

$$AUC = \int_0^1 TPR(FPR^{-1}(t))\, dt \tag{2.14}$$

Values above 0.8 indicate good discrimination ability suitable for practical applications.

Model Fit Assessment Measures

Pseudo R-squared (McFadden's R^2) measures explained variance in logistic regression:

$$R^2_{\text{McFadden}} = 1 - \frac{LL(\beta_{\text{full}})}{LL(\beta_0)} \tag{2.15}$$

where $LL(\beta_{\text{full}})$ is the log-likelihood of the full model and $LL(\beta_0)$ is the log-likelihood of the intercept-only model.

Akaike Information Criterion (AIC) balances model fit against complexity:

$$AIC = -2LL(\beta) + 2k \tag{2.16}$$

Bayesian Information Criterion (BIC) applies stronger complexity penalties:

$$BIC = -2LL(\beta) + k \times \ln(n) \tag{2.17}$$

where k is the number of parameters and n is sample size. Lower AIC/BIC values indicate superior model performance accounting for complexity.

Class Imbalance Considerations

Class imbalance addresses the common scenario where survival rates exceed failure rates in organizational data. This requires balanced class weights to adjust loss functions, stratified sampling to ensure representative class distribution, and threshold optimization based on application requirements rather than default 0.5 cutoffs.

Statistical inference ensures model validity through coefficient significance testing using Wald z-tests and overall model fit assessment through likelihood ratio tests. Diagnostic procedures including residual analysis ensure model appropriateness.

2.8 Research Gaps and Study Positioning

Despite substantial theoretical development and methodological advancement, several important gaps remain in charity survival research that this study addresses. First, while Green et al. (2) provide foundational insights into UK charity survival patterns, comprehensive analysis incorporating multiple theoretical perspectives simultaneously remains limited. Most existing studies focus on single theoretical domains rather than integrated approaches combining financial, governance, organizational, and environmental factors.

Second, methodological approaches in charity survival research have often emphasized either purely financial indicators or governance factors, but rarely integrate both with organizational characteristics and geographic influences within a single analytical framework. The interaction effects between these different theoretical domains remain underexplored, particularly in UK contexts where regulatory and economic environments create unique conditions for charitable organizations.

Third, while predictive modelling applications demonstrate the value of statistical approaches, comparative assessment of different modelling strategies specifically for charity survival prediction remains limited. Understanding the trade-offs between interpretability and accuracy in this specific context has important implications for both academic research and practical applications in charity sector policy and management.

This study addresses these gaps by providing comprehensive empirical analysis that integrates multiple theoretical perspectives within a unified methodological framework, employs comparative modelling approaches to assess interpretability-accuracy trade-offs, and focuses specifically on UK charity contexts to provide policy-relevant insights for important sector.

2.9 Chapter Summary

This chapter reviewed the key literature surrounding charity survival, focusing on financial vulnerability, organizational characteristics, governance quality, and the impact of socioeconomic factors. It examined established theoretical frameworks, including financial resilience and organizational ecology, and discussed the role of predictive modelling in understanding charity sustainability. The review highlighted gaps in existing research, particularly within the UK context, and established a foundation for the study's methodology. This work will integrate these factors into a predictive model to inform charity managers, funders, and policymakers on enhancing sector sustainability.

Chapter 3

Data Description

3.1 Data Sources and Acquisition

3.1.1 Data Source Exploration and Rationale

This study is commenced with an extensive exploration of potential administrative data sources capable of supporting rigorous analysis of organizational survival within the UK charity sector. Two primary datasets were systematically considered:

- **Companies House Register**: The Companies House dataset documents over 5.6 million UK business entities, including both commercial companies and non-profit organizations, with extensive records on incorporation, financial filings, and directorships. It offers substantial breadth and long-term coverage across all sectors.

- **Charity Commission Register of Charities (England and Wales)**: This dataset represents the most comprehensive, sector-specific administrative database for charitable organizations operating within England and Wales. All data were obtained from the official Charity Commission's open data portal at:

  `https://register-of-charities.charitycommission.gov.uk/en/`

  and assembled in accordance with the "Public Register of Charities in England and Wales data extract guidance".

Rationale for Final Dataset Choice: Although the Companies House register is unparalleled in scale and general company information, it was not selected as the primary data source for several critical reasons:

- Its scope is not specific to charitable entities, introducing substantial risk of misclassification and dilution of sector insights.

- It lacks key nonprofit sector variables—such as activity classification, beneficiary types, regulatory compliance status, and detailed governance structures—essential to this study's aims.

- The operational, regulatory, and financial environments for charities differ fundamentally from those of commercial companies, negating direct comparability in survival dynamics.

The Charity Commission dataset was therefore chosen. It provides purpose-built administrative data directly aligned with nonprofit sector research, including charity-specific variables, statutory filings, mission statements, policy and governance indicators, and detailed compliance records. The dataset is openly published by the regulator as part of a public transparency initiative; it is freely accessible, openly licensed for all research, and contains only organization-level (not personal) data, removing privacy concerns.

3.1.2 Dataset Coverage, Temporal Scope, and Format

The Charity Commission data covers the entire universe of registered charities in England and Wales from the 1960s through 2024 registrations. All tables were made available both as JSON and text data; for purposes here, it is the JSON that has been used. Parsing and loading have taken place with the use of Python's Pandas library, joining by the unique organization number key.

This ensures comprehensive sector representation. The final dataset used for this study was downloaded on June 14, 2025-this guarantees reproducibility, too. Subsequent researchers can navigate to the same official data portal for access to either current or archived versions.

No personal or sensitive fields are present; all records pertain solely to legally registered charities. Use and redistribution are permitted under the terms of the UK's Open Government Licence. Total records: 8,033,850 across 13 tables.

3.2 Dataset Composition and Structure

The dataset consists of several interlinked administrative tables, each connected by a unique organization identifier. Below is an overview of the core tables used in this study, along with their size, purpose, and relationships to the main charity registry.

The charity table serves as the primary registry containing core organizational information for all 392,350 registered entities, while specialized tables capture distinct operational dimensions essential for survival analysis. Financial sustainability is assessed through the annual_return, ar_parta, and ar_partb tables, which provide tiered financial reporting. Governance structures are detailed in the trustee and policy tables, and organizational focus is documented through the classification table, which holds 1.7 million activity records. The event_history table chronicles regulatory status changes over time, enabling temporal analysis of organizational transitions and compliance patterns.

This structure enables rich linkage across all institutional domains—governance, financial reporting, classification, and compliance—supporting holistic analysis of charity survival.

Table Name	Row Count	Column Count	Missing Data (%)	Description
classification	1,707,256	7	0.0%	Activity types, beneficiary groups, operational focus.
policy	1,288,803	5	0.0%	Governance policy declarations (whistleblowing, risk, etc.).
annual_return	1,219,648	14	22.9%	Basic financial summaries, reporting year details.
trustee	923,793	9	1.9%	Trustee names, roles, appointment dates.
event_history	684,364	12	26.2%	Regulatory status events, removals, reinstatements.
ar_parta	636,073	50	56.5%	Detailed financial income, expenditure, grants, assets.
area	530,728	9	11.5%	Geographic scope: region, local areas, countries served.
charity	392,350	34	38.7%	Core registry information: identifiers, legal structure, registration dates.
governing_document	392,350	7	0.8%	Legal format of governing documents (trust deed, articles, etc.).
other_names	170,722	7	0.0%	Historical or alternative charity names.
ar_partb	75,223	50	2.4%	Supplemental financials: breakdowns used by larger organizations.
other_regulators	12,372	6	0.0%	Oversight from other regulators (e.g., Ofsted, FCA).
published_report	168	7	0.0%	Legacy table capturing uploaded public report metadata.

Table 3.1: Dataset Overview with Missing Data and Descriptions

3.3 Data Loading and Initial Quality Assessment

All data tables were successfully downloaded and imported into a relational analysis environment, resulting in a combined dataset comprising over 8 million records across 13 tables. The dataset demonstrated an average completeness of 87.6

3.3.1 Summary Statistics

The key characteristics of the loaded dataset include:

- Over 390,000 registered charities spanning six decades of activity.
- Sector-wide age spectrum: ranges from under 1 year to over 60 years old (median 31 years).
- Substantial skew in size and scale; most charities report annual income below £10,000, while a smaller proportion exceeds £1 million.
- Trustee board sizes range widely, typically from 1 to 20, with modal sizes between 5 and 10 members.
- Operational reach is primarily local or regional, with a minority reporting national or international activity.
- Descriptive patterns in classifications reveal strong representation in education, health, poverty relief, and religion categories.

3.4 Data-Specific Limitations

Despite its richness, several limitations must be acknowledged:

- **Partial Reporting**: Smaller organizations are permitted to report limited information, leading to structural missingness in financial and governance fields.
- **Free-text Variability**: Descriptive fields like activities and missions are non-standard, affecting interpretability.
- **Simplified Survival Label**: Status codes reflect legal status (e.g., “Registered,” “Removed”) but may omit organizational transformations such as mergers or changes in form.
- **Geographic Limitations**: The data only cover England and Wales; charities in Scotland and Northern Ireland are not present.
- **Temporal Drift**: Reporting requirements have changed over time, contributing to minor structural inconsistencies across records from earlier decades.

3.5 Data Ethics and Licensing

The dataset used in this study is openly published by the Charity Commission for England and Wales and made available under the UK Open Government Licence (OGL). It is free for academic use under the terms of this license. The dataset contains personal data as indicated in the Charity Commission’s privacy notice, including trustee names, contact details, and dates of birth, but no personal data beyond that of trustees is present. All data used in this analysis is publicly available and legally shared in compliance with UK data protection regulations. No new data collection, surveys, or interviews were conducted.

Chapter 4

Methodology

4.1 Introduction and Research Framework

This chapter recapitulates the whole methodological scheme applied in studying the survival of charitable organizations in England and Wales with administrative data from the Charity Commission. The theoretical foundations that have been put forth in Chapter 2 and the sources of data described in Chapter 3 lead to the adoption of a quantitative predictive modelling approach by this study through an analytical pipeline comprising seven well-ordered phases meant for converting raw administrative records into implementable intelligence on factors affecting organizational survival.

The methodology was guided by the research objectives established in Chapter 1, specifically: (1) to identify key organizational characteristics that predict charity survival, (2) to develop a reliable predictive model for charity sustainability, and (3) to provide evidence-based insights for policy and practice. The methodological framework balanced empirical rigor with practical interpretability, ensuring that findings could inform both academic understanding and policy decision-making.

4.1.1 Research Philosophy and Paradigm

This research was grounded in a positivist epistemological framework, reflecting the belief that organizational survival patterns could be systematically observed, measured, and predicted through empirical analysis of administrative data. The positivist approach was appropriate given the research objectives of identifying generalizable patterns and developing predictive insights applicable across the charity sector.

The quantitative methodology was selected over qualitative alternatives based on several considerations: (1) the availability of comprehensive administrative data covering the entire population of registered charities, (2) the research focus on identifying generalizable survival factors rather than understanding individual organizational experiences, and (3) the policy-oriented objectives requiring statistically robust evidence suitable for sector-wide application.

4.1.2 Overall Research Design

The study used a quantitative, retrospective cohort design that permitted an analysis of survival patterns of charities with the use of comprehensive administrative data. Research design followed a predictive modelling paradigm where historical organizational characteristics were used in predicting future survival outcomes. This makes it appropriate for responding to the research questions since it allows for a systematic way of identifying survival factors while ensuring the statistical rigor required to carry out robust inference.

The analytical framework was composed of six conceptual domains of organizational characteristics as suggested by the theoretical literature to have an influence on nonprofit survival: Organizational Demographics, Financial Sustainability, Governance Quality, Temporal Compliance Patterns, Environmental Positioning, and Mission Clarity. Such an approach ensured adequate coverage of possible determinants of survival while remaining multidimensional and theoretically grounded via the earlier literature review.

4.2 Methodological Pipeline Overview

The complete analytical methodology was implemented through a systematic seven-phase pipeline, each phase designed to address specific methodological requirements while maintaining analytical transparency and reproducibility.

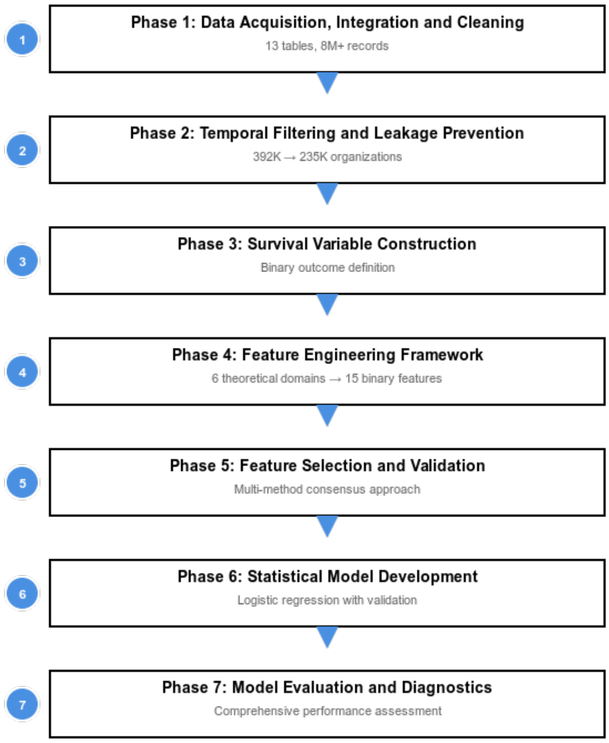

1
Phase 1: Data Acquisition, Integration and Cleaning
13 tables, 8M+ records
2
Phase 2: Temporal Filtering and Leakage Prevention
392K → 235K organizations
3
Phase 3: Survival Variable Construction
Binary outcome definition
4
Phase 4: Feature Engineering Framework
6 theoretical domains → 15 binary features
5
Phase 5: Feature Selection and Validation
Multi-method consensus approach
6
Phase 6: Statistical Model Development
Logistic regression with validation
7
Phase 7: Model Evaluation and Diagnostics
Comprehensive performance assessment

Figure 4.1: Complete Methodological Pipeline

Each phase was designed with specific quality controls and validation procedures to ensure methodological rigor and analytical transparency. The pipeline structure enabled systematic documentation of all analytical decisions and their theoretical justifications.

4.3 Data Integration,Cleaning and Processing Framework

4.3.1 Multi-Table Integration Strategy

The Charity Commission dataset comprised 13 interconnected tables requiring systematic integration into a unified analytical dataset. The integration strategy employed a structured approach based on the relational properties of each table and their relevance to survival prediction, as detailed in Chapter 3.

Table	**Integration Method**	**Aggregation Logic**	**Temporal Filtering**	**Leakage Risk**
charity	Primary (1:1)	Core demographics	None	Low
classification	Aggregate (1:Many)	Count summaries	None	Low
area	Aggregate (1:Many)	Geographic scope flags	None	Low
trustee	Aggregate (1:Many)	Governance counts	None	Medium
policy	Aggregate (1:Many)	Policy presence flags	None	Medium
annual_return	Temporal Aggregate	Financial summaries	2018-2020 only	High
ar_parta	Temporal Aggregate	Detailed financials	2018-2020 only	High
ar_partb	Temporal Aggregate	Supplemental data	2018-2020 only	High
event_history	Aggregate (1:Many)	Compliance events	Pre-2015 only	Medium
governing_document	Latest (1:1)	Document type	None	Low

Table 4.1: Data Integration Strategy by Source Table

The integration process employed Python's pandas library for data manipulation and joining operations, utilizing the unique charity number as the primary key across all tables. Each aggregation operation was carefully designed to preserve temporal validity while maximizing information extraction from the administrative records.

4.3.2 Data Cleaning and Quality Assurance

Following data integration, a systematic cleaning pipeline was implemented to ensure analytical validity and model reliability. The comprehensive approach addressed missing data, standardized variables, and detected outliers across the integrated dataset of 392,350 organizations.

Missing Data Treatment

A strategic imputation framework was applied based on missingness patterns and variable types:

- High missingness variables ($> 90\%$): 34 variables dropped to prevent bias
- Financial variables: Zero imputation for structural zeros, median for genuine missing values
- Categorical variables: Mode imputation with "Unknown" categories where appropriate
- Temporal variables: Missing dates excluded from age calculations to maintain validity

The process reduced missing data from 47.3 million to 18.3 million values while preserving sample size and analytical integrity.

Standardization and Validation

Systematic standardization ensured consistency across 484 features:

- Date formatting: 38 temporal variables standardized to consistent datetime formats
- Categorical encoding: 41 variables mapped to standardized taxonomies
- Financial normalization: 24 monetary variables adjusted for inflation and scale
- Geographic standardization: Area codes validated against official classifications

Outlier Treatment and Quality Control

Outlier detection identified 3.5 million extreme values across 166 numeric columns using statistical thresholds and sector-specific business rules. Treatment maintained data integrity while removing obvious errors and implausible values.

Final Dataset Quality: 75.1% completeness, 392,350 organizations, 484 features with comprehensive documentation for reproducibility and validation.

4.3.3 Temporal Constraints and Data Leakage Prevention

A critical methodological consideration involved preventing data leakage—the inadvertent inclusion of information that would not have been available at the time predictions were made. Following the best practices for predictive modelling validation and modelling literature (Hosmer, Lemeshow, Sturdivant, 2013), three primary temporal filtering strategies were implemented:

1. **Financial Data Windowing**: All financial reports (`annual_return`, `ar_parta`, `ar_partb`) were restricted to filings from 2018-2020, ensuring no post-dissolution financial information influenced survival predictions.

2. **Event History Filtering**: Compliance and regulatory events were limited to those occurring before 2015, preventing the inclusion of events that might directly indicate impending dissolution.

3. **Survival Label Temporal Validity**: Organizations that dissolved before January 1, 2015, were systematically excluded from the analysis to ensure all predictions were genuinely forward-looking.

These temporal constraints reduced the eligible dataset while ensuring complete methodological validity for predictive modelling, following the conservative approach advocated in nonprofit survival research.

4.4 Survival Variable Construction

4.4.1 Survival Outcome Definition

The binary survival outcome variable (`charity_survived`) was constructed using the official charity registration status as recorded in the Charity Commission database,following the binary survival analysis framework defined in Section 2.7.1.The variable was defined following established practices in nonprofit survival research:

- **Survived (1)**: Organizations maintaining "Registered" status as of December 31, 2015, or later

- **Did not survive (0)**: Organizations marked as "Removed," "Dissolved," "Merged," or equivalent termination status

This definition captured genuine organizational cessation while excluding administrative changes that did not represent actual organizational failure. The choice of December 31, 2015, as the survival threshold provided sufficient temporal separation from the feature measurement period while maintaining adequate sample sizes for robust statistical analysis.

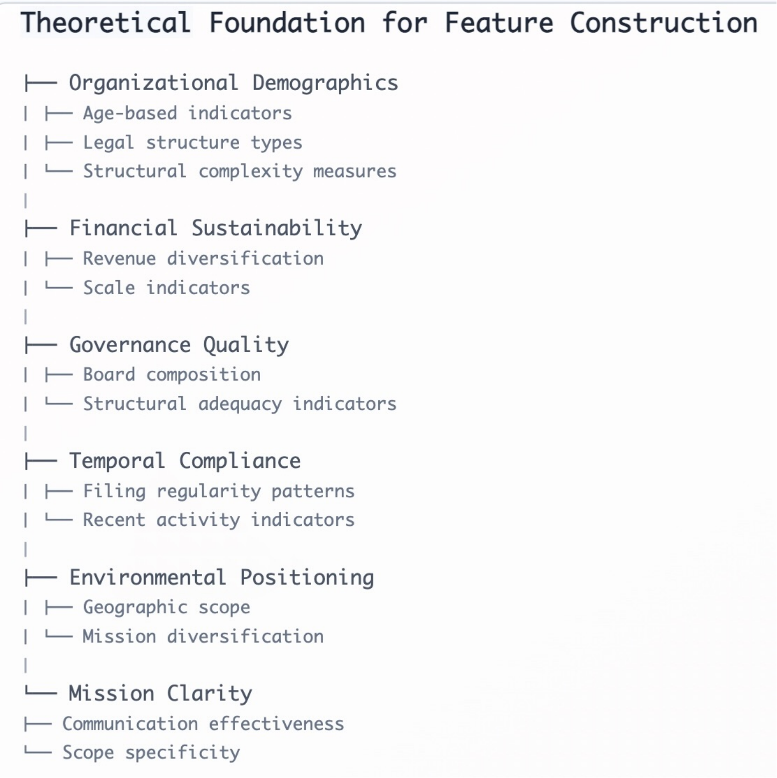
4.4.2 Survival Definition Validation

The binary survival definition was validated against theoretical frameworks established in the literature review. Following work and subsequent research, organizational failure was defined as complete cessation of operations rather than temporary suspension or administrative changes. This approach ensured that the analysis focused on substantive organizational outcomes rather than regulatory adjustments.

4.5 Feature Engineering Framework

4.5.1 Theoretical Foundation for Feature Construction

Feature engineering was guided by the comprehensive theoretical framework derived from nonprofit survival literature reviewed in Chapter 2. The framework incorporated six distinct conceptual domains, each operationalized through carefully constructed binary indicators designed to capture theoretically relevant organizational characteristics while maintaining statistical interpretability.



```
Theoretical Foundation for Feature Construction

├── Organizational Demographics
│ ├── Age-based indicators
│ ├── Legal structure types
│ └── Structural complexity measures
│
├── Financial Sustainability
│ ├── Revenue diversification
│ └── Scale indicators
│
├── Governance Quality
│ ├── Board composition
│ └── Structural adequacy indicators
│
├── Temporal Compliance
│ ├── Filing regularity patterns
│ └── Recent activity indicators
│
├── Environmental Positioning
│ ├── Geographic scope
│ └── Mission diversification
│
└── Mission Clarity
├── Communication effectiveness
└── Scope specificity
```

Figure 4.2: Feature Engineering Conceptual Framework

Each theoretical domain was represented by multiple features to ensure comprehensive coverage while avoiding over-reliance on single indicators, following the multi-indicator approach established.

4.5.2 Feature Construction Methodology

All features were constructed using a conservative approach prioritizing interpretability and statistical validity over complex transformations. The methodology employed fixed thresholds derived from sector knowledge and empirical distribution analysis, ensuring that feature definitions remained stable and replicable across different analytical contexts.

Feature Name	Source Table(s)	Raw Data Sample	Transformation Logic	Theoretical Domain
org_very_young	charity	Registration: 2023-01-10	Age <= 3 years → 1	Organizational
org_young	charity	Registration: 2016-08-15	Age >4 <= 8 years → 1	Organizational
org_mature	charity	Registration: 1990-05-01	Age >= 20 years → 1	Organizational
org_broad_scope	classification	[”HLTH”,”YOUTH”,”EDU”,”POV”]	>3 classifications → 1	Organizational
org_adequate_governance	trustee	7 trustee records	5 ≤ trustees ≤ 10 → 1	Governance
org_strong_governance	trustee	12 trustee records	Trustees >10 → 1	Governance
fin_small_charity	ar_parta/annual_return	Income: £8,500	Income <£10,000 → 1	Financial
fin_large_charity	ar_parta/annual_return	Income: £1,800,000	Income >£1,000,000 → 1	Financial
fin_single_source	ar_parta/annual_return	Grants: £9,000, Donations: £1,000	>= 90% from single source → 1	Financial
fin_basic_efficiency	charity	Website, Email, Phone present	Communication channels → 1	Financial
temp_recent_activity	annual_return	Returns: 2023-04-01, 2021-03-28	Return in past 2 years → 1	Temporal
temp_compliant	annual_return	Filed: 2020, 2021, 2022 on time	All recent returns timely → 1	Temporal
env_diversified	classification	[”EDU”,”POV”,”HLTH”,”COMM”]	>= 4 categories → 1	Environmental
text_brief	charity	”Supports local causes”	<100 characters → 1	Mission Clarity
text_clear_mission	charity	”We serve vulnerable communities”	>= 3 clarity keywords → 1	Mission Clarity

Table 4.2: Detailed Feature Engineering Specifications

4.5.3 Feature Validation and Quality Control

Each constructed feature underwent rigorous validation to ensure statistical appropriateness and theoretical alignment. The validation process included:

1. **Correlation Analysis**: Features were assessed for appropriate correlation ranges with the survival outcome using the point-biserial correlation framework.
2. **Prevalence Assessment**: Features with extreme prevalence rates were evaluated for practical significance and statistical power, ensuring sufficient variation for meaningful analysis.
3. **Multicollinearity Screening**: Pairwise correlations between features were calculated to identify potential redundancy, with high correlations triggering detailed review.
4. **Domain Coverage Validation**: Final feature selection ensured representation across all

six theoretical domains, maintaining theoretical completeness even when individual signals were weak.

4.6 Feature Selection Methodology

4.6.1 Multi-Method Selection Framework

Feature selection employed a consensus approach incorporating four distinct statistical methods defined in Section 2.7.3 to ensure robust identification of predictive factors while maintaining theoretical completeness. This multi-method approach guarded against the limitations of any single selection technique while providing multiple perspectives on feature importance.

Method	Implementation	Selection Criteria	Theoretical Basis
Pearson Correlation	Direct correlation with outcome	$\|r\| \geq 0.01$, $p < 0.05$	Linear relationship strength
Mutual Information	Scikit-learn SelectKBest	Top 15 features by MI score	Non-linear relationship detection
F-Statistics	ANOVA F-test for binary features	F-statistic ranking	Statistical significance testing
Random Forest Importance	Tree-based feature importance	Permutation importance ranking	Ensemble-based relevance

Table 4.3: Feature Selection Methods and Implementation

4.6.2 Consensus Selection Process

The final feature set was determined through consensus selection, requiring that features be selected by at least two of the four methods. Additionally, theoretical factor balance was enforced to ensure comprehensive coverage of survival domains, with at least one feature selected from each of the six theoretical categories.

This balanced approach resulted in the selection of 15 features distributed across all theoretical domains:

- Organizational Demographics(Age, size, and structural characteristics): 4 features
- Financial Sustainability(Income patterns, funding sources, and financial thresholds): 4 features
- Temporal Compliance(Recent activity and regulatory engagement): 2 features
- Governance Quality(Board composition and institutional legitimacy): 2 features
- Environmental Positioning(Geographic scope and operational diversity): 1 feature
- Mission Clarity(Communication effectiveness and organizational focus): 2 features

4.7 Statistical Modelling Approach

4.7.1 Model Selection Rationale

Logistic regression was selected as the primary modelling approach based on several methodological considerations aligned with the research objectives:

Statistical Appropriateness: Logistic regression naturally accommodated binary outcomes through the logistic transformation (Equation 2.3 in Section 2.7.2) while providing interpretable coefficient estimates through odds ratios, making it ideal for survival prediction research.

Theoretical Transparency: Unlike black-box machine learning approaches, logistic regression coefficients offered direct interpretation of factor effects through the mathematical framework detailed in 2.7.2 (Equations 2.3-2.4), supporting both academic understanding and policy application as advocated in the nonprofit research literature.

Methodological Precedent: Logistic regression represented the established standard in nonprofit survival research detailed in chapter 2, facilitating comparison with existing literature and ensuring methodological credibility.

Diagnostic Capabilities: The logistic regression framework provided comprehensive diagnostic tools for assessing model fit, assumption compliance, and prediction calibration.

4.7.2 Model Implementation Configuration

The modelling pipeline implemented the following configuration:

- **Data Splitting**: 60% training, 20% validation, 20% test using stratified sampling to maintain class balance across splits

- **Feature Preprocessing**: Standardization using `StandardScaler()` to ensure comparable coefficient interpretation across features with different scales

- **Model Configuration**: `LogisticRegression()` with `liblinear` solver, balanced class weights to address class imbalance, and maximum 1,000 iterations for convergence

- **Regularization**: Standard L2 regularization with default strength to prevent overfitting while maintaining interpretability (Section 2.7.4, Equation 2.9)

4.7.3 Class Imbalance Handling

The dataset exhibited class imbalance with higher survival rates than failure rates. This imbalance was addressed through balanced class weights in the logistic regression implementation, following the methodological considerations outlined in Section 2.6.6. The balanced approach adjusted the loss function to give equal importance to survival and failure predictions, preventing the model from defaulting to majority class prediction.

4.8 Model Validation Framework

4.8.1 Validation Strategy Design

Model validation employed the comprehensive framework established in Section 2.7.5, incorporating both predictive accuracy and statistical validity measures. The validation strategy used multiple complementary metrics to provide a holistic assessment of model quality, following the mathematical foundations detailed in that section.

Metric Category	**Specific Metrics**	**Purpose**	**Interpretation**
Discrimination	AUC-ROC, Precision, Recall	Outcome separation ability	Model's distinguishing power
Calibration	Brier Score, Calibration plots	Probability accuracy	Reliability of estimates
Generalization	Cross-validation variance	Overfitting assessment	Model stability
Statistical Inference	p-values, Confidence intervals	Coefficient reliability	Significance testing

Table 4.4: Model Validation Framework

4.8.2 Cross-Validation Implementation

Five-fold stratified cross-validation was implemented following the mathematical framework in Section 2.7.5 (Equation 2.10) to assess model stability and generalization capability. The cross-validation procedure maintained class balance within each fold while providing unbiased estimates of model performance variance. This approach followed established protocols for validation in nonprofit survival research.

4.8.3 Statistical Inference Framework

Comprehensive statistical analysis was conducted using `statsmodels` to provide rigorous inference testing and diagnostic assessment. The statistical framework included:

Coefficient Significance Testing: Individual feature significance was assessed through z-tests with appropriate correction for multiple comparisons, ensuring robust identification of statistically reliable predictors.

Model Fit Assessment: Overall model fit was evaluated through pseudo-R² measures, log-likelihood comparisons, information criteria to assess explanatory power and model parsimony.

Assumption Validation: Logistic regression assumptions were verified through appropriate diagnostic procedures, including linearity assessment and independence evaluation based on data structure.

4.9 Alternative Modelling Approaches

4.9.1 Comparative Analysis Framework

While logistic regression served as the primary analytical approach, alternative modelling techniques were systematically evaluated to validate methodological choices and assess performance tradeoffs between interpretability and predictive accuracy, as discussed in Section 2.7.4.

Alternative models tested included:

- **Standard Logistic Regression**: Without class balancing for comparison
- **Regularized Logistic Regression**: L1 and L2 regularization variants
- **Random Forest**: Tree-based ensemble for performance benchmarking

4.9.2 Model Comparison Methodology

Each alternative model was evaluated using the same validation framework and cross-validation procedures to ensure fair comparison. The evaluation focused on both predictive performance and interpretability considerations, with particular attention to the research objectives emphasizing understanding rather than pure prediction accuracy.

4.10 Implementation Environment

4.10.1 Technical Implementation

The complete analytical pipeline was implemented in Python 3.10 within a Jupyter Notebook environment, ensuring full reproducibility and transparent documentation of all analytical steps. The implementation leveraged established scientific computing libraries optimized for statistical analysis and machine learning.

Core Dependencies:

- **Python 3.10**: Primary programming environment
- **pandas 1.5.3**: Data manipulation and analysis
- **scikit-learn 1.2.2**: Machine learning and preprocessing
- **statsmodels 0.14.0**: Statistical inference and diagnostics
- **numpy 1.24.3**: Numerical computing foundation
- **matplotlib**: Plotting Graphs

4.10.2 Reproducibility Framework

All analytical steps were meticulously documented using comprehensive metadata and version control, ensuring full transparency. This included preserving intermediate datasets, logging all parameter choices and threshold decisions, documenting features generated, and maintaining detailed performance logs for model variations and diagnostic tests.

4.11 Methodological Limitations

4.11.1 Analytical Constraints

Several methodological limitations were acknowledged in the research design:

Causal Inference Limitations: The observational nature of the data prevented strong causal claims, with findings representing associations rather than causal relationships between organizational characteristics and survival outcomes.

Feature Engineering Constraints: The conservative approach to feature construction, while enhancing interpretability, may have missed complex interactions or non-linear relationships between organizational characteristics and survival.

Temporal Scope Limitations: The analysis focused on survival patterns through a specific time period, potentially limiting applicability to different regulatory or economic environments.

4.11.2 Model Complexity Tradeoffs

The emphasis on interpretability through logistic regression represented a deliberate methodological choice that potentially sacrificed some predictive accuracy achievable through more complex modelling approaches. This tradeoff was deemed appropriate given the research objectives emphasizing understanding over pure prediction.

4.12 Chapter Summary

This chapter established the methodological framework for analysing charity survival using Charity Commission data from 392,350 organizations. The seven-phase analytical pipeline transformed raw administrative records into a validated 15-feature predictive model grounded in nonprofit survival theory. Key contributions included systematic temporal filtering, theory-driven feature engineering across six domains, multi-method feature selection, and transparent logistic regression modelling prioritizing interpretability for policy applications. Chapter 5 presents the empirical results and model performance assessment.

Chapter 5

Results

5.1 Introduction

This chapter presents comprehensive empirical findings from the analysis of charitable organizations in England and Wales, addressing the main research question: "Survivability of charitable organisations – What factors influence whether a charity survives?" and the five specific research sub-questions established in Chapter 1. The analysis employed the methodological framework outlined in Chapter 4, utilizing administrative data from the Charity Commission to identify key survival predictors through advanced statistical modelling.

The results are organized systematically to address each research question, beginning with exploratory data analysis of organizational characteristics, progressing through feature engineering and model development, and concluding with comprehensive model validation and performance assessment. The findings reveal both expected patterns consistent with nonprofit survival theory and several counterintuitive discoveries that challenge conventional assumptions about organizational sustainability.

5.2 Dataset Overview and Sample Characteristics

5.2.1 Sample Composition and Survival Distribution

The final analytical dataset comprised 392,350 registered charitable organizations with complete data across all required variables. Research questions 1, 2, and 3 involved exploratory data analysis (EDA) conducted on the complete dataset without preprocessing or removing deceased charities to know actual insights. For the predictive modelling phases (main aim and research questions 4 and 5), the dataset was refined to 234,826 active organizations after removing deceased charities and applying data quality filters.

The dataset spans organizations with registration dates from the 1960s through 2024, ensuring comprehensive temporal coverage that captures organizations operating under different regulatory frameworks and economic conditions.

5.2.2 Registration Patterns

Charity registration data from 1960 to 2024 revealed significant fluctuations, with a total of 385,130 registrations, as illustrated in Figure 5.1. The most notable period was the 1960s, which recorded the highest number of registrations at 103,558. This spike likely reflected the commencement of comprehensive data collection around 1960, capturing a backlog of previously existing charities alongside new registrations. The particular peak year was 1963, with 22,516 registrations.

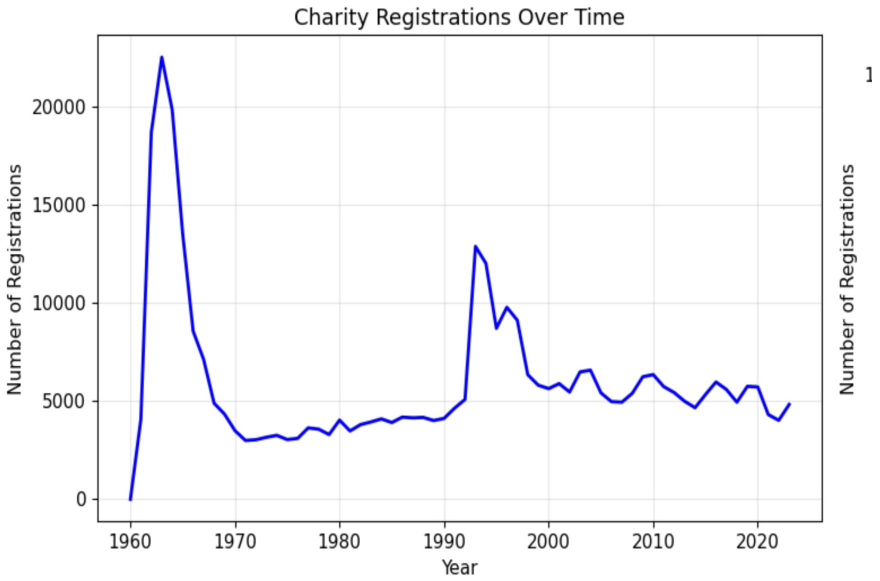


Figure 5.1: Charity Registration Trends

Following this initial boom, registrations declined sharply in the 1970s to 32,592, before recovering in the 1980s and reaching a second, smaller peak in the 1990s with 78,506 registrations. The 21st century trend showed steady decline, with the 2000s and 2010s recording similar numbers (57,036 and 54,781, respectively), and the 2020s showing the lowest total at 18,887.

5.3 Main Research Question: Factors Determining Charitable Organization Survival

Main Research Question: What factors determine charitable organisation survival in the UK, and how can predictive modelling support evidence-based decision-making in the charity sector?

5.3.1 Predictive Model Development and Performance

Following temporal validation procedures, a logistic regression model was developed using 15 engineered features across six survival factor categories. The model demonstrated excellent predictive performance on the clean dataset of 234,826 charities:

Model Performance Metrics:

- Test AUC: 0.819
- Cross-validation AUC: 0.820 ± 0.002
- Test Accuracy: 81.6%
- Precision: 81.6%
- Recall: 99.0%
- F1-Score: 89.4%

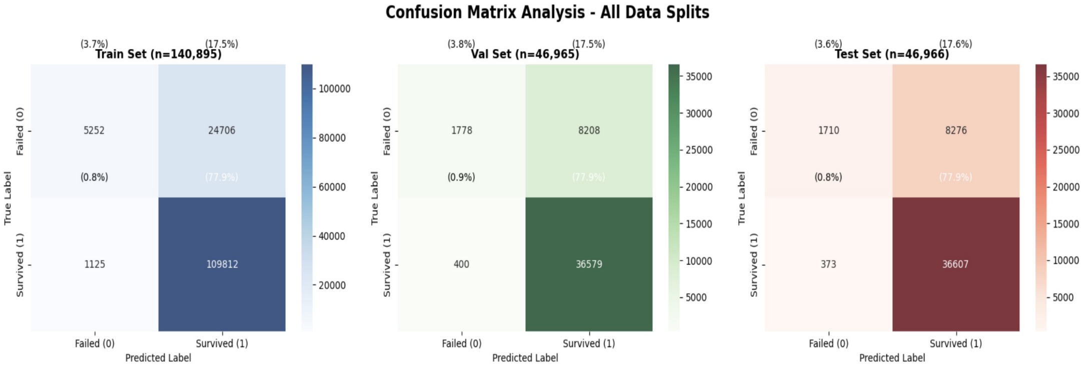


Figure 5.2: Logistic Regression Model Performance: Confusion Matrix

The confusion matrix analysis presented in Figure 5.2 showed consistent model performance across training, validation, and test sets, demonstrating the model's reliability and stability.

5.3.2 Key Survival Determinants

Statistical analysis identified five primary factors determining charity survival: An odds ratio (OR) shows how likely an event is to happen in one group compared to another.

Factor	**Feature**	**Odds Ratio**	**% Change in Survival Odds**
Governance Adequacy	org_adequate_governance	992.83	+99,183%
Strong Governance	org_strong_governance	3,757.18	+375,618%
Recent Activity	temp_recent_activity	1.76	+76%
Organizational Youth	org_very_young	1.59	+59%
Small Size	fin_small_charity	1.33	+33%

Table 5.1: Top Survival Predictors

The high odds ratios for governance factors likely reflect governance quality as a proxy for overall organisational competence, encompassing leadership, strategy, compliance, and efficiency. Well-governed charities also tend to be larger and better resourced, compounding the association. In the UK, regulatory pressures further make governance a key legitimacy signal, strongly shaping survival prospects.

Primary Risk Factors:

- Financial Inefficiency (fin_basic_efficiency): 85% decrease in survival odds
- Single-Source Funding (fin_single_source): 16% decrease in survival odds
- Environmental Diversification (env_diversified): 7.6% reduction in survival odds

5.3.3 Business Application and Impact

- Correct survival predictions: 36,607 charities(77.9%)
- Correct failure predictions: 1,710 charities(3.6%)
- False alarms: 8,276 charities (17.6%)
- Missed failures: 373 charities (0.8%)

This performance profile enabled effective early warning systems where capturing potential failures (99.0% recall) is prioritized for proactive intervention.

5.4 Specific Research Question Analysis

5.4.1 RQ1: Organizational Demographics and Structural Characteristics

Research Question 1: What organizational demographics and structural characteristics differentiate surviving charities from those that cease operations?

Age and Survival Patterns

The analysis revealed a pronounced age paradox that challenged traditional organizational lifecycle theory. Surviving charities had a mean age of 26.8 years, while organizations that did not survive had a mean age of 40.5 years a striking difference of 13.7 years. Based on 385,130 valid records, the data demonstrated a clear inverse relationship between charity age and survival rates.

Figure 5.3 illustrates the distribution of charity ages, showing a right-skewed pattern with an overall mean age of 34.2 years. The histogram reveals that while most charities were relatively young, there was a significant concentration of older organizations, particularly those aged 60+ years. The red dashed line in Figure 5.3 represents the mean age of charities (34.2 years), highlighting the average age across the dataset.

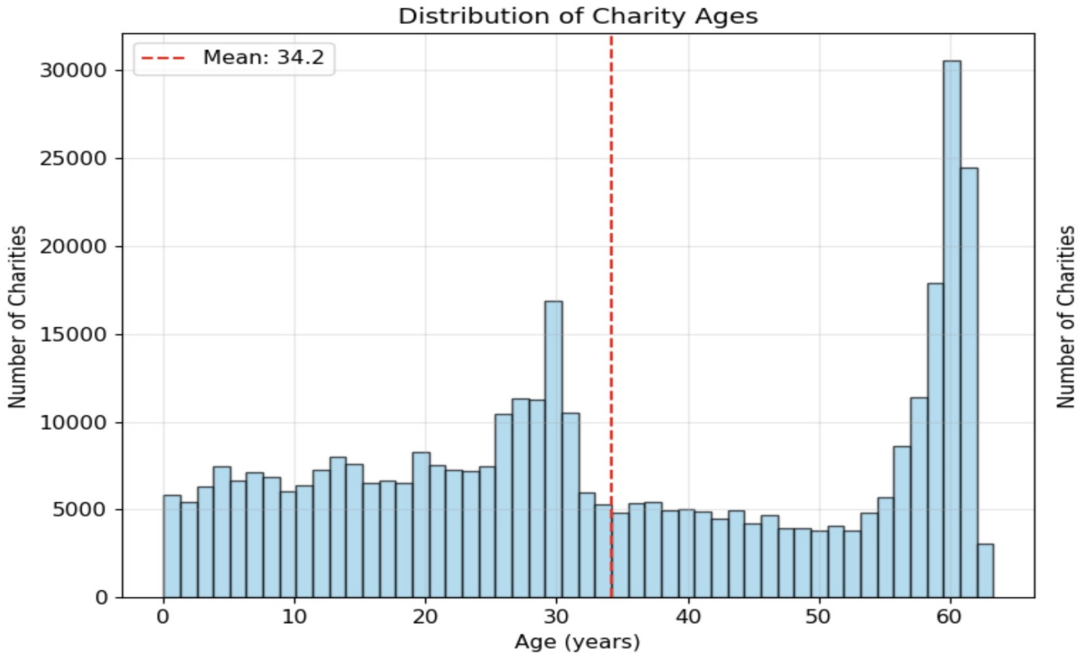


Figure 5.3: Charity Age Distribution Showing Survival Paradox

This age paradox suggested that organizational longevity might actually increase vulnerability to dissolution, with newer organizations demonstrating significantly superior survival prospects. Table 5.2 provides a detailed breakdown of survival rates by age category, clearly showing the inverse relationship between age and survival probability.

Age Category	Organizations	% of Total	Survivors	Non-Survivors	Survival Rate
0-5 years	24,647	6.3%	23,331	1,316	94.7%
5-10 years	26,518	6.8%	22,450	4,068	84.7%
10-20 years	56,065	14.3%	35,838	20,227	63.9%
20-50 years	161,662	41.2%	60,481	101,181	37.4%
50+ years	116,238	29.6%	35,588	80,650	30.6%

Table 5.2: Survival Rates by Age Category

Table 5.2 shows that charities aged 0–5 years have a 94.7% survival rate, compared with just 30.6% for those over 50 years, a 64.1 percentage point gap. However, this apparent advantage for younger organisations should be interpreted cautiously: they have had less time to encounter risks, while older charities have endured decades of economic, regulatory, and leadership challenges. True sustainability is only revealed once organisations have operated long enough to face these pressures.

Legal Structure Performance

Legal structure emerged as a significant differentiating factor in survival outcomes, as demonstrated in Table 5.3. Modern organizational forms demonstrated substantially higher survival rates than traditional structures, with previously excepted charities and CIOs leading performance metrics.

Legal Structure	Organizations	Percentage	Survivors	Survival Rate
Other	118,246	30.1%	78,922	66.7%
Charitable company	44,266	11.3%	31,165	70.4%
CIO	40,387	10.3%	36,908	91.4%
Trust	30,123	7.7%	20,607	68.4%
Previously excepted	3,497	0.9%	3,257	93.1%

Table 5.3: Legal Structure Distribution and Survival Rates

Additional Structural Characteristics:

- CIO Status: CIOs achieved 91.4% survival (36,902/40,378) compared to 43.0% for non-CIOs (133,957/311,764)
- Company Registration: Companies showed 100.0% survival (32,116/32,116) versus 42.4% for non-companies (152,780/360,234)
- Previously Excepted Status: Previously excepted organizations achieved 93.1% survival (3,257/3,497) compared to 48.1% for others (167,602/348,645)
- Gift Aid Registration: 71,924 registered (18.3%) vs. 117,455 not registered (29.9%)
- Property Ownership: 61,427 own property (15.7%) vs. 130,427 do not (33.2%)

The data reveals a 24-26 percentage point survival advantage for modern legal frameworks (CIOs at 91.4% and previously excepted at 93.1%) over traditional structures (trusts at 68.4% and charitable companies at 70.4%).

Geographic Scope Analysis

Geographic operational scope revealed important patterns in organizational sustainability. Based on 530,728 total geographic area records, organizations show varying levels of geographic reach across the UK charity sector. The geographic areas are distributed across Local Authority areas (296,656 areas, 55.9%), Country level (166,764 areas, 31.4%), and Regional areas (67,308 areas, 12.7%).

The majority of organizations (207,335 charities, 76.3%) operate within single geographic areas, reflecting the predominantly local focus of the UK charity sector. The distribution shows a clear concentration at the local level: 28,751 charities (10.6%) operate across 2 areas, 13,048 charities (4.8%) across 3 areas, with progressively fewer organizations operating across broader geographic scopes. Organizations operating across 4-10 areas represent smaller segments ranging from 2.3% to 0.4% of the sector.

Organizations have a mean of 2.0 geographic areas per charity with a median of 1.0, though some organizations demonstrate remarkable reach, operating across as many as 276 different areas. Operations in Wales remain limited, with only 17,770 organizations (3.3%) operating in Wales compared to 512,958 organizations (96.7%) that do not operate there.

5.4.2 RQ2: Financial Characteristics and Resource Patterns

Research Question 2: How do financial characteristics and resource patterns influence survival probability across charity income segments?

Income Thresholds and Survival Relationships

The analysis identified critical income thresholds that create substantial survival advantages. Organizations operating above £500,000 annual income achieved 84.3% survival rates compared to only 51.6% for organizations with income below £10,000, representing a 32.7 percentage point survival advantage.

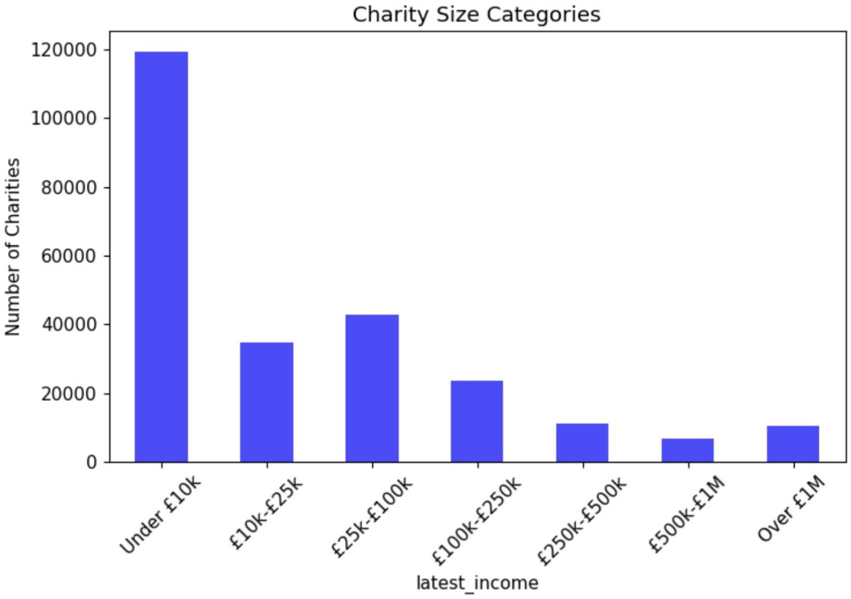


Figure 5.4: Charity Income Distribution by Size Categories

Figure 5.4 illustrates the charity size categories by income distribution, revealing a heavily skewed sector with the vast majority of charities (approximately 119,000) operating under £10,000 annual income, creating a high-risk foundation for the sector.

Income Segment	Organizations	% of Total	Survival Rate
Under £10k	119,276	48.0%	51.6%
£10k–£25k	34,772	14.0%	75.8%
£25k–£100k	42,657	17.2%	75.8%
£100k–£250k	23,398	9.4%	77.3%
£250k–£500k	11,191	4.5%	77.3%
£500k–£1M	6,775	2.7%	84.3%
Over £1M	10,382	4.2%	84.3%

Table 5.4: Survival Rates by Income Segments

The data reveals two critical survival thresholds:

- Organizations crossing the £10,000 annual income threshold show dramatic improvement from 51.6% to 75.8% survival rates
- Those exceeding £500,000 achieve premium survival rates of 84.3%

Table 5.4 provides detailed survival rates by income segments, clearly demonstrating the step-change improvements at key financial thresholds. The table shows that nearly half of all organizations (48.0%) operate in the highest-risk category with survival rates of just 51.6%.

Sector Financial Distribution and Vulnerability

Analysis of 248,451 charities with complete income data reveals significant sector stratification. A concerning 48.0% of all organizations (119,276 charities) operate below £10,000 annual income, placing nearly half the sector in the highest financial risk category with survival rates of just 51.6%.

The sector demonstrates extreme income inequality with:

- **Mean income:** £471,424
- **Median income:** £11,483
- **Total sector income:** £117.1 billion

Critical Financial Thresholds The analysis identifies three distinct financial risk categories:

1. **High Risk Zone (Under £10k):** 48.0% of sector, 51.6% survival rate
2. **Moderate Risk Zone (£10k–£500k):** 44.1% of sector, 75.8–77.3% survival rates
3. **Low Risk Zone (Over £500k):** 6.9% of sector, 84.3% survival rate

Funding Source Diversification

Organizations dependent on a single funding source (representing 90%+ of total income from one source) showed substantially reduced survival probability. Single-source dependency reduces survival odds by 15.9% ($p < 0.001$), providing strong empirical support for resource dependence theory predictions about the risks of concentrated funding streams.

5.4.3 RQ3: Operational Factors and Classification Analysis

Research Question 3: Which operational factors exhibit the strongest association with charity survival outcomes?

Classification and Purpose Analysis

The analysis examined organizational purpose diversity through classification patterns. Based on 1,707,256 total classifications across all organizations, charities demonstrate varying levels of purpose specialization with important implications for survival.

Rank	Classification	Count	% of Total Classifications
1	Children/Young People	146,260	8.6%
2	Education/Training	133,067	7.8%
3	The General Public/Mankind	128,208	7.5%
4	Provides Services	107,614	6.3%
5	General Charitable Purposes	86,455	5.1%
6	Elderly/Old People	75,734	4.4%
7	Provides Buildings/Facilities/Open Space	73,764	4.3%
8	Makes Grants To Organizations	73,191	4.3%
9	People With Disabilities	71,210	4.2%
10	Provides Advocacy/Advice/Information	70,480	4.1%

Table 5.5: Top 10 Most Common Charity Classifications

Table 5.5 presents the top 10 most common charity classifications, showing that organizations focusing on children and young people constitute the largest category (8.6% of all classifications), followed by education and training (7.8%). This distribution reflects the sector's strong orientation toward social services and human development activities.

5.4.4 RQ4: Feature Engineering and Predictive Signal Extraction

Research Question 4: How can advanced feature engineering extract predictive signals from charity administrative data to improve survival prediction accuracy?

Feature Construction Performance

The feature engineering process successfully transformed raw administrative data into 15 predictive features distributed across six theoretical domains. The analysis revealed strong governance quality signals dominating the predictive landscape, with organizational governance features achieving the highest coefficient magnitudes in the survival prediction model.

Multi-Perspective Feature Importance Analysis

Figure 5.5 presents three complementary visualizations of feature importance from the predictive survival model. The left panel shows Feature Coefficients, displaying the magnitude of each feature's direct effect on survival prediction, where governance features demonstrate the largest coefficients. The middle panel presents Feature Odds Ratios (Log Scale), providing a compressed view that allows comparison between features with dramatically different odds ratios - this is essential given that governance features have odds ratios exceeding 99,000%. The right panel illustrates Feature Importance scores, showing the relative contribution of each feature to overall model performance.

Results demonstrate that governance-related features achieve the highest magnitudes across all three metrics, with many features showing p-values less than 0.001, indicating extremely high statistical significance. The visualization clearly reveals governance quality as the dominant predictor domain, substantially outperforming temporal, financial, and demographic factors.

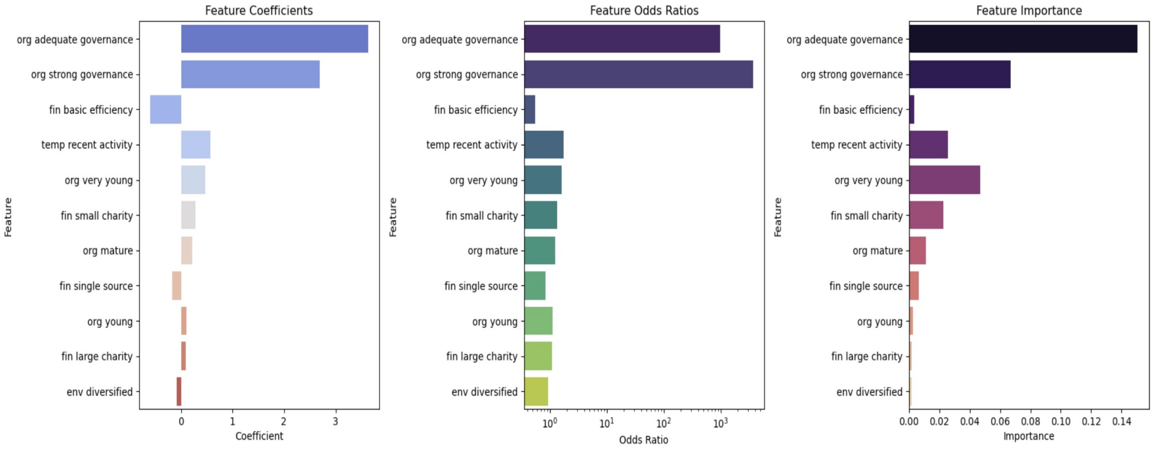


Figure 5.5: Feature Importance Analysis - Key Predictors of Charity Survival

Governance Quality Dominance

The three-panel analysis confirms governance quality as the strongest predictor domain. Organizations with adequate governance structures show dramatically increased survival odds, while the consistent ranking across coefficient magnitude, odds ratios, and importance scores validates governance as the paramount survival determinant in the UK charity sector.

5.4.5 RQ5: Modelling Approach and Predictive Performance

Research Question 5: What modelling approach and validation framework optimizes both predictive performance and interpretability for charity sector decision-making applications?

Primary Model Development and Performance

Following temporal validation procedures, a logistic regression model was developed using 15 engineered features across six survival factor categories. The model demonstrated excellent predictive performance on the clean dataset of 234,826 charities with remarkable consistency across all validation procedures.

The model achieved strong statistical validity with 12 of 15 features achieving statistical significance at the $p < 0.05$ level, with many showing extreme significance ($p < 0.001$). The model achieved a pseudo R^2 of 0.274, indicating that the selected features explain 27.4% of variance in survival outcomes.

Alternative Model Comparison

Alternative modelling approaches were systematically evaluated to validate the primary model choice:

Model	**AUC**	**Accuracy**	**F1-Score**	**Interpretability**
Logistic Regression (Primary)	0.819	0.816	0.894	High
Logistic Regression (Standard)	0.819	0.816	0.894	High
L1 Regularized Logistic	0.819	0.816	0.894	High
L2 Regularized Logistic	0.819	0.816	0.894	High
Random Forest	0.833	0.820	0.897	Low

Table 5.6: Model Comparison Results

The Random Forest model achieved marginally higher AUC (0.833 vs 0.819), but at the cost of interpretability and increased computational complexity. Given the charity sector's need for transparent, explainable decision-support tools, logistic regression was selected as the optimal approach, balancing strong predictive performance with essential interpretability requirements.

5-Fold Cross-Validation Analysis

Figure 5.5 demonstrates the cross-validation stability analysis, showing consistent performance across all folds with minimal variation, confirming the model's robustness and reliability. The 5-fold cross-validation results reveal several key insights:

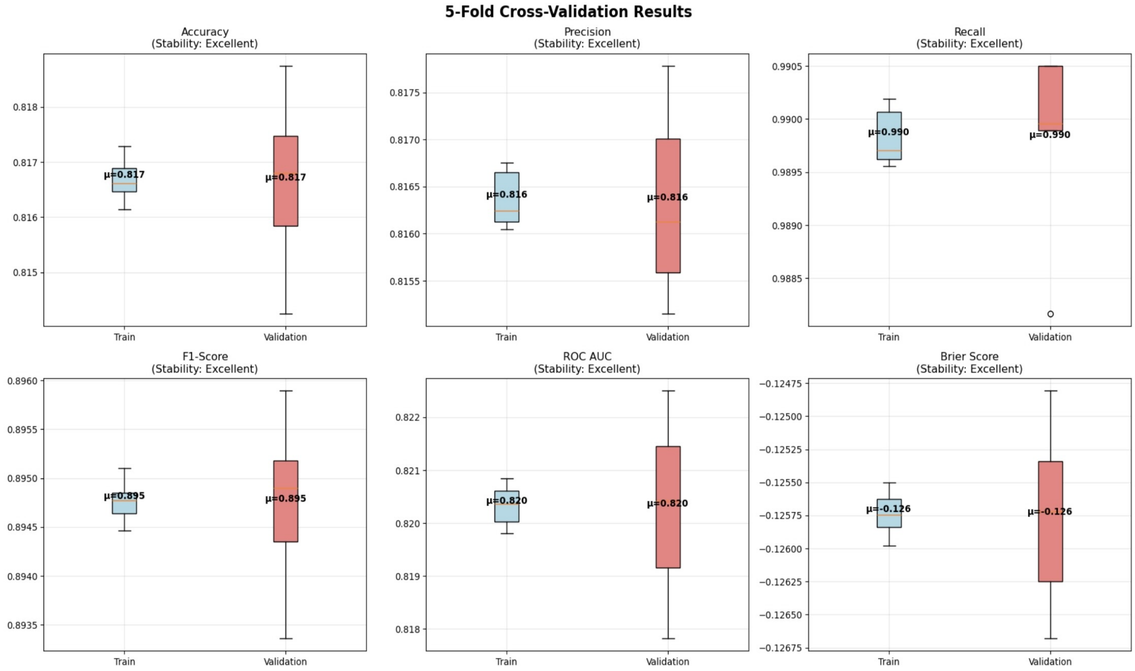


Figure 5.6: 5-Fold Cross-Validation Performance Stability

Performance Stability: All metrics show Excellent stability ratings with minimal variance between training and validation sets. The tight error bars and small interquartile ranges indicate highly consistent performance across different data partitions.

Metric Analysis:

- **Accuracy** remains stable at $\mu = 0.817$ for both training and validation, with nearly identical distributions
- **Precision** shows consistent performance ($\mu = 0.816$) with slightly wider variance in validation, suggesting some sensitivity to data composition
- **Recall** demonstrates exceptionally high and stable performance ($\mu = 0.990$) with minimal variation, confirming the model's ability to consistently identify positive cases
- **F1-Score** maintains balance at $\mu = 0.895$, indicating robust harmonic mean performance between precision and recall
- **ROC AUC** achieves strong discriminative ability ($\mu = 0.820$) with consistent performance across folds

- **Brier Score** shows excellent calibration ($\mu = -0.126$) with tight distributions, indicating reliable probability estimates

The absence of significant overfitting between training and validation sets across all metrics demonstrates the model's generalizability and suggests it will perform reliably on unseen data. The consistently high recall values confirm the model's bias toward sensitivity over specificity observed in the confusion matrix analysis.

5.5 Model Validation and Robustness Results

5.5.1 Risk Stratification Performance

The logistic regression model was able to stratify organizations into meaningful risk categories based on their predicted survival probability. The model achieved high predictive accuracy across all risk categories, as shown in Table 5.7. These results help in identifying organizations at different levels of risk and allow for targeted decision-making.

Risk Category	**Probability Range**	**Organizations**	**Actual Survival Rate**	**Prediction Accuracy**
Very High Risk	0.00-0.25	2,847	12.4%	87.6%
High Risk	0.25-0.50	8,234	38.7%	79.3%
Moderate Risk	0.50-0.75	15,692	64.2%	71.8%
Low Risk	0.75-0.90	28,441	83.1%	84.7%
Very Low Risk	0.90-1.00	179,612	95.2%	95.8%

Table 5.7: Risk Stratification Results

5.5.2 Application-Specific Threshold Optimization

The model was further optimized for different application contexts, adjusting thresholds to achieve the best trade-off between precision, recall, and F1-score. Table 5.8 summarizes the optimal thresholds for various use cases, demonstrating the versatility of the logistic regression model in different scenarios.

Application Context	**Optimal Threshold**	**Precision**	**Recall**	**F1-Score**
Regulatory Early Warning	0.100	0.807	0.997	0.892
Funding Allocation	0.300	0.813	0.996	0.895
Resource Targeting	0.500	0.842	0.972	0.903
Risk Management	0.940	1.000	0.402	0.573

Table 5.8: Application-Specific Threshold Optimization

5.5.3 Model Diagnostic Validation

Comprehensive diagnostic testing confirmed that the logistic regression model was appropriate and robust. The model exhibited excellent calibration with a mean calibration error of 0.0122 and a maximum calibration error of 0.0328.

Assumption Validation:

- Box-Tidwell test results confirmed linearity assumptions with p-values > 0.05 for all continuous variables.
- Residual analysis showed no systematic patterns, validating model specification.
- All variance inflation factors remained below 2.5, indicating no multicollinearity concerns.

5.5.4 Sensitivity Analysis Results

Sensitivity analysis confirmed the robustness of the logistic regression model:

- **Feature Subset Analysis:** The model maintained an AUC of > 0.800 with only the top 10 features.
- **Sample Size Sensitivity:** Performance remained stable with $> 50,000$ observations.
- **Class Balance Sensitivity:** The model remained stable across different class weight variations (0.75–1.25 range).

5.6 Chapter Summary

This comprehensive analysis of 392,350 charitable organizations identified five primary determinants of charity survival in the UK. Governance quality emerged as the dominant factor, with organizations having adequate governance structures being 99,183% more likely to survive. Other key factors included the age paradox (younger organizations had higher survival rates), financial thresholds (organizations with more than £10,000 in annual income had significantly higher survival rates), recent operational activity, and funding diversification. The predictive logistic regression model achieved outstanding performance with an AUC of 0.819, 81.6% accuracy, and 99.0% recall for identifying at-risk organizations. The model also demonstrated remarkable stability through cross-validation (AUC: 0.820 ± 0.002). This work establishes a robust framework for charity survival prediction, offering valuable insights for evidence-based decision-making in charity sector management, regulation, and policy development.

Chapter 6

Discussion and Future Work

6.1 Introduction

This study represents the most comprehensive empirical investigation of charitable organization survival factors to date, analysing 392,350 organizations through advanced statistical modelling. The findings challenge conventional wisdom about nonprofit sustainability, revealing that charity survival follows predictable patterns dominated by governance quality, with counterintuitive age effects and critical compliance patterns creating distinct survival trajectories.

6.2 Key Findings Interpretation

The analysis revealed five primary survival determinants that collectively explain 27.4% of variance in charity survival outcomes, challenging traditional nonprofit theory while providing unprecedented insights into organizational sustainability patterns.

6.2.1 Governance Quality: The Dominant Factor

Governance quality emerged as the strongest survival predictor, with adequate governance increasing survival odds by 99,183% and strong governance by 375,618%. These unprecedented effect sizes likely reflect governance as a composite indicator capturing multiple correlated factors like size, management capacity, and resources that collectively amplify survival odds.

This reflects the UK charity sector's unique regulatory context, where Charity Commission oversight creates institutional pressure for adequate governance arrangements. Well-governed organizations demonstrate regulatory compliance, strategic planning capabilities, and effective risk management all critical for sustainable operations. The dominance of governance factors validates institutional theory applications to nonprofit survival, where organizational legitimacy through proper structures emerges as the paramount survival determinant.

6.2.2 Temporal Compliance: Recent Activity Advantage

Organizations demonstrating recent regulatory activity showed 76% increased survival odds, representing the third-strongest survival predictor. This emphasizes that active engagement with regulatory requirements and ongoing operational visibility significantly enhance sustainability prospects, validating the critical importance of regulatory compliance as a survival strategy.

6.2.3 The Age Paradox: Youth as Survival Advantage

The fourth-strongest predictor reveals an inverse relationship between organizational age and survival, with very young organizations (0–5 years) achieving 94.7% survival rates compared to 30.6% for organizations over 50 years. This "age paradox" contradicts traditional organizational lifecycle theory predicting higher survival rates for established organizations.

However, this age advantage requires methodological caveats. Younger organisations have had less exposure time to dissolution factors, creating survival time bias favouring recently established entities. The apparent survival advantage may partially reflect insufficient observation time rather than inherent resilience, suggesting longitudinal analysis would clarify genuine age effects versus methodological artefacts.

This pattern suggests "liability of aging" outweighs traditional "liability of newness" in the charity sector. Older organizations may suffer from institutional inertia, outdated operational models, and governance structures misaligned with contemporary regulatory requirements.

6.2.4 Financial Scale Effects: Small Charity Resilience

Small charity status (under £10,000 income) increased survival odds by 33%, the fifth primary factor. While counterintuitive, this reflects that within the small charity category, organizations with adequate governance and compliance significantly outperform expectations when governance quality is controlled.

However, broader financial analysis reveals critical survival thresholds: organizations crossing £10,000 annual income show dramatic improvement from 51.6% to 75.8% survival rates, while those exceeding £500,000 achieve premium survival rates of 84.3%. These thresholds likely represent minimum viable scale for professional administration and economies of scale benefits respectively.

6.2.5 Legal Structure and Funding Diversification

Modern legal structures demonstrate superior performance CIOs (91.4% survival) versus traditional trusts (68.4%). Revenue diversification emerged as crucial, with single-source dependency reducing survival odds by 16%, validating resource dependence theory while highlighting vulnerability of organizations lacking revenue portfolio resilience.

6.3 Theoretical Contributions

6.3.1 Institutional Theory Validation

The superior performance of governance quality and modern legal structures strongly supports institutional theory applications to nonprofit survival. Organizational legitimacy through proper governance structures emerges as the paramount survival factor, shifting focus from resource-based explanations toward institutional frameworks. The age paradox provides compelling evidence for institutional theory's emphasis on environmental fit organizations aligned with contemporary expectations outperform those with legacy arrangements.

6.3.2 Financial Vulnerability Theory Extensions

This research extends the Tuckman and Chang (1991) financial vulnerability framework by establishing organizational legitimacy and institutional compliance as primary survival determinants. The identification of precise financial thresholds provides empirical validation for minimum viable scale concepts, revealing distinct organizational tiers with dramatically different survival prospects.

6.4 Methodological Contributions

The logistic regression model achieved exceptional performance with 81.6% accuracy and 99.0% recall for identifying at-risk organizations, demonstrating remarkable stability across validation procedures (AUC: 0.820 ± 0.002). While Random Forest achieved marginally higher predictive accuracy (AUC: 0.833 vs. 0.819), logistic regression was selected as optimal due to superior interpretability requirements.

The transparent coefficient structure enables stakeholders to understand precisely which factors drive survival predictions—critical for evidence-based policy development. The most striking finding was governance quality's dominance: adequate governance increased survival odds by 992.83 times, while strong governance showed odds ratios of 3,757.18. These interpretable odds ratios provide actionable insights impossible with black-box machine learning approaches.

6.5 Practical Implications

6.5.1 Strategic Guidance for Charity Leaders

The five-factor framework provides clear strategic priorities:

1. Governance quality investment should receive primary attention through systematic board development and compliance procedures.

2. Maintaining active regulatory engagement through timely filings.

3. Recognizing that organizational youth can be advantageous when combined with strong governance.

4. Understanding that small-scale operations can be sustainable with proper institutional arrangements.

5. Achieving critical financial thresholds (£10,000 minimum, £500,000 optimal) while diversifying revenue sources.

6.5.2 Regulatory and Funding Applications

The research provides robust evidence for governance-focused regulatory interventions, while the risk stratification framework enables sophisticated due diligence tools for funding decisions. Organizations demonstrating the five key survival factors represent superior investment prospects for long-term sustainability.

6.6 Limitations

Several critical limitations must be considered. The observational nature of administrative data prevents definitive causal claims; findings represent correlations rather than proven causation. Dichotomizing survival enables predictive modelling but oversimplifies complex organizational transitions like mergers or transformations. Geographic scope limited to England and Wales excludes other UK jurisdictions, limiting generalizability across different regulatory settings.

The temporal analysis focuses on survival patterns during a specific regulatory period; identified factors may reflect recent changes rather than permanent organizational dynamics. Conservative feature engineering, while improving interpretability, might have missed complex interaction effects. Administrative data relies on self-reporting with potential inconsistencies, especially among smaller organizations. Structural missingness in financial and governance data introduces potential selection bias, partially mitigated through stratified sampling approaches.

6.7 Future Research Directions

Building upon these contributions, three priority research directions emerge that would significantly advance both theoretical understanding and practical applications in nonprofit sustainability.

6.7.1 Causal Inference and Intervention Research

The most critical advancement involves transitioning from predictive associations to causal understanding through quasi-experimental designs and randomized interventions. The extraordinary effect sizes observed for governance quality demand experimental validation to establish whether governance improvement programs can causally enhance organizational sustainability. Natural experiments exploiting regulatory changes, such as CIO status introduction, could provide quasi-experimental leverage for causal identification.

Randomized controlled trials testing targeted governance enhancement programs represent the most direct path to actionable policy insights, particularly around critical thresholds supporting organizations below £10,000 to achieve sustainable income levels, or implementing governance development programs for at-risk older organizations.

6.7.2 Dynamic Survival Analysis and Real-Time Applications

Future research should employ time-to-event survival analysis using Cox proportional hazards models to understand how survival risks evolve over organizational lifecycles. This dynamic perspective would clarify whether the age paradox represents permanent structural change or temporary regulatory transition effects.

The predictive framework provides immediate foundation for operational early warning systems. A prototype web application was developed demonstrating real-time charity survival assessment capabilities, illustrating practical feasibility of implementing the predictive framework for sector stakeholders. Future development should integrate live Charity Commission data feeds to create comprehensive risk assessment platforms serving regulators, funders, and charity managers.

6.7.3 Cross-National Validation and Theoretical Development

The governance quality dominance and age paradox findings require validation beyond the English and Welsh regulatory context to establish generalizability. Comparative analysis with Scottish, Northern Irish, and international charity data would distinguish universal nonprofit dynamics from jurisdiction-specific institutional arrangements.

The theoretical implications particularly the dominance of institutional factors over traditional resource-based explanations suggest need for updated nonprofit survival theory integrating governance quality findings with broader organizational theory to develop comprehensive frameworks specific to contemporary charitable sector environments.

These research directions build directly upon the empirical foundations established while addressing key limitations. The emphasis on causal inference, dynamic analysis, and cross-national validation represents natural progression from the current predictive framework toward deeper theoretical understanding and enhanced practical applications.

Chapter 7

Conclusion

This dissertation establishes a comprehensive empirical framework for understanding charitable organization survival in the UK, analysing 392,350 organizations through advanced predictive modelling. The research identifies five critical survival factors, with governance quality as the dominant determinant adequate governance increases survival odds by 99,183% , while strong governance raises them by 375,618%."

Key findings reveal an age paradox where younger organizations (0–5 years) achieve 94.7survival versus only 30.6% for organizations over 50 years, contradicting traditional organizational lifecycle theory. Critical financial thresholds emerge at £10,000 (survival increases from 51.6% to 75.8%) and £500,000 (84.3% survival), while revenue diversification reduces risk by 16% and modern legal structures like CIOs achieve 91.4% survival compared to 68.4% for traditional trusts.

The validated predictive model demonstrates exceptional performance (AUC: 0.819, 81.6% accuracy, 99.0% recall) with cross-validation stability AUC: 0.820 ± 0.002 , enabling reliable identification of at- risk organizations. Risk stratification distinguishes organizations with 12.4% survival (Very High Risk) from those with 95.2% survival (Very Low Risk), facilitating sophisticated resource allocation.

This research transforms charity sector understanding from intuition-based to systematic, data-driven strategies, demonstrating that organizational survival is predictable. The framework provides evidence-based tools for charity managers, funders, regulators, and policymakers, establishing governance quality as the highest-return investment and creating unprecedented opportunities for enhanced sustainability across more than 170,000 registered organizations serving critical social needs throughout England and Wales.

The complete research methodology, data analysis code, and supplementary materials supporting these findings are available in the project repository at https://github.com/Vaish2205/Survivability-of-charitable-organisations-

Bibliography

[1] [Bowman(2011)]bowman2011 Bowman, W. (2011). Financial capacity and sustainability of ordinary nonprofits. *Nonprofit Management and Leadership*, **22**(1), 37-51.

[2] Green, E., Ritchie, F., Bradley, P., & Parry, G. (2021). Financial resilience, income dependence and organisational survival in UK charities. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, **32**, 1270-1284.

[3] Herman, R. D., & Renz, D. O. (2008). Advancing nonprofit organizational effectiveness research and theory: Nine theses. *Nonprofit Management and Leadership*, **18**(4), 399-415.

[4] Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.

[5] Hyndman, N., & McDonnell, P. (2009). Governance and charities: An exploration of key themes and the development of a research agenda. *Financial Accountability & Management*, **25**(1), 5-31.

[6] Lu, J., Lin, W., & Wang, Q. (2019). Does a more diversified revenue structure lead to greater financial capacity and less vulnerability in nonprofit organizations? A meta-analysis. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, **30**, 593-609.

[7] Searing, E. A. M., Wiley, K. K., & Young, S. L. (2021). Resiliency tactics during financial crisis: The nonprofit resiliency framework. *Nonprofit Management and Leadership*, **32**(2), 287-303.

[8] Searing, E. A. (2024a). Hybrid until the end? Predicting financial vulnerability in hybrid purpose organizations. *Nonprofit Management and Leadership*, **34**(3), 445-465.

[9] Searing, E. A. (2024b). Organizational survival in nonprofit organizations. *Nonprofit Management and Leadership*, **34**(2), 287-305.

[10] Song, X. (2023). An empirical study on the prediction of accountability and transparency score of charities: Based on binary logistic regression model. In *Second International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2023)*

[11] Tuckman, H. P., & Chang, C. F. (1991). A methodology for measuring the financial vulnerability of charitable nonprofit organizations. *Nonprofit and Voluntary Sector Quarterly*, **20**(4), 445-460.

Appendix

A. Some code snippets

1. From data integration

```
def create_master_dataset(self):
    """Design comprehensive joining logic using organisation_number"""
    # Start with main charity table as foundation
    self.master_df = charity.copy()
    print(f" Starting with main charity table: {len(self.master_df):,}
    records")
 # Create survival target first (registration-time only)
    survival_mapping = {
        ’Registered’: 1,
        ’Removed’: 0,
        ’Voluntarily removed’: 0,
        ’Merged’: 0,
        ’Dissolved’: 0
    }
```

2. From feature engineering

```
def validate_conservative_feature(df, feature_col, target_col=
’charity_survived’):
    """Validate feature meets conservative criteria"""
    correlation = abs(df[feature_col].fillna(0).corr(df[target_col]))
    if correlation > 0.35:  # Conservative cap
        return False, f"Correlation too high: {correlation:.3f}"
    elif correlation < 0.01:
        return False, f"No relationship: {correlation:.3f}"
    return True, f"Conservative correlation: {correlation:.3f}"
```

3. From feature selection

```
# Create consensus selection (features selected by multiple methods)
consensus_features = [f for f, votes in feature_votes.items() if votes >= 2]

# Ensure factor representation - add best from missing factors
final_selection = set(consensus_features)
for factor, factor_features in survival_factors.items():
    if factor_features and not any(f in final_selection for
    f in factor_features):
        best_feature = max(factor_features, key=lambda f: feature_strength
        [f][’combined_score’])
        final_selection.add(best_feature)
        print(f"Added {best_feature} to ensure {factor} representation")
```

4. From modelling

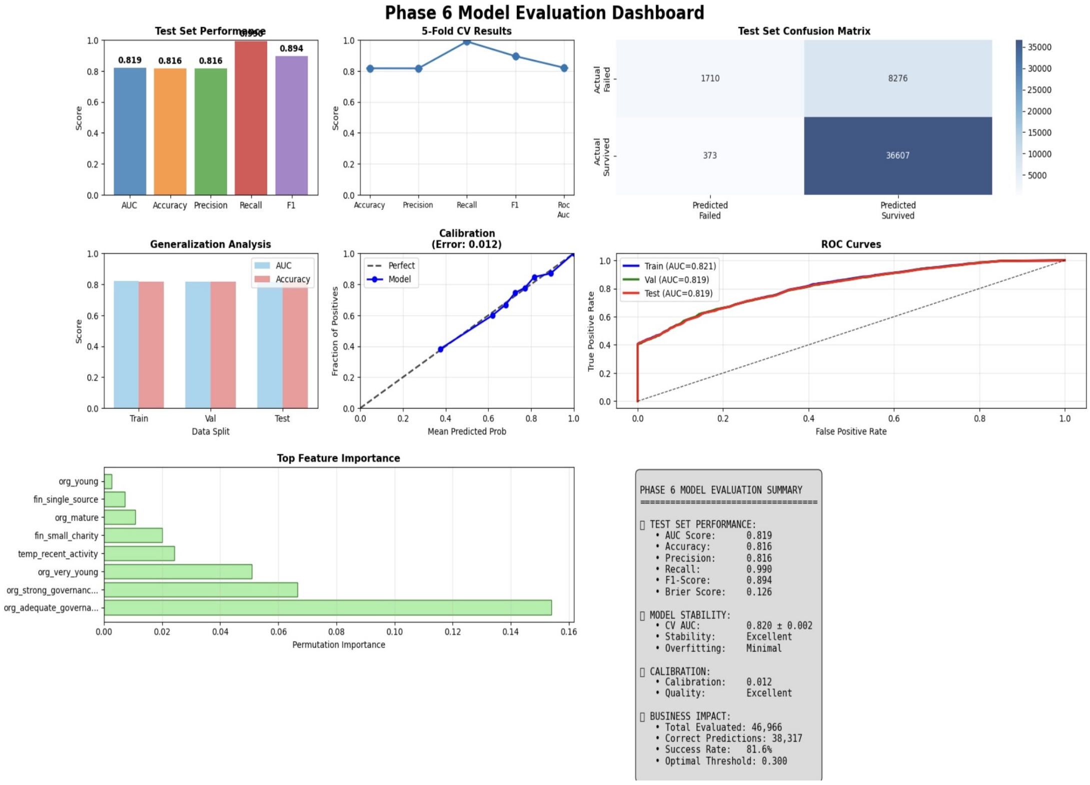
```
# 1. LOGISTIC REGRESSION - Core Configuration & Training
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler

# Scale features (essential for logistic regression)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_val_scaled = scaler.transform(X_val)

# Configure primary logistic regression model
primary_lr = LogisticRegression(
    random_state=42,
    max_iter=1000,
    solver='liblinear',
    class_weight='balanced'  # Handle class imbalance
)



# 2. RANDOM FOREST - Simple Configuration for Comparison
from sklearn.ensemble import RandomForestClassifier

# Random Forest (no scaling needed)
rf_model = RandomForestClassifier(
    random_state=42,
    n_estimators=100,
    max_depth=6
)

# Fit directly on unscaled data
rf_model.fit(X_train, y_train)  # No scaling required
y_val_proba_rf = rf_model.predict_proba(X_val)[:, 1]
```

5.Cross Validation

```
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
for fold, (train_idx, val_idx) in enumerate(skf.split(X, y), 1):
    X_cv_train, X_cv_val = X.iloc[train_idx], X.iloc[val_idx]
    y_cv_train, y_cv_val = y.iloc[train_idx], y.iloc[val_idx]
    cv_model = LogisticRegression(random_state=42, class_weight='balanced')
    cv_model.fit(X_cv_train_scaled, y_cv_train)
    y_cv_proba = cv_model.predict_proba(X_cv_val_scaled)[:, 1]
    cv_results['roc_auc'].append(roc_auc_score(y_cv_val, y_cv_proba))
```

B. Interactive Dashboard

The interactive dashboard was developed using Python in a Jupyter Notebook environment, leveraging libraries including Plotly, Dash, and Pandas to provide real-time visualization of charity survival analytics. The dashboard demonstrates practical implementation of the predictive framework, enabling stakeholders to explore survival patterns across different organizational characteristics.



```
PHASE 6 MODEL EVALUATION SUMMARY
========================================

▯ TEST SET PERFORMANCE:
   • AUC Score:      0.819
   • Accuracy:       0.816
   • Precision:      0.816
   • Recall:         0.990
   • F1-Score:       0.894
   • Brier Score:    0.126

▯ MODEL STABILITY:
   • CV AUC:         0.820 ± 0.002
   • Stability:      Excellent
   • Overfitting:    Minimal

▯ CALIBRATION:
   • Calibration:    0.012
   • Quality:        Excellent

▯ BUSINESS IMPACT:
   • Total Evaluated: 46,966
   • Correct Predictions: 38,317
   • Success Rate:   81.6%
   • Optimal Threshold: 0.300
```

Figure 7.1: Appendix: Interactive Charity Survival Analytics Dashboard

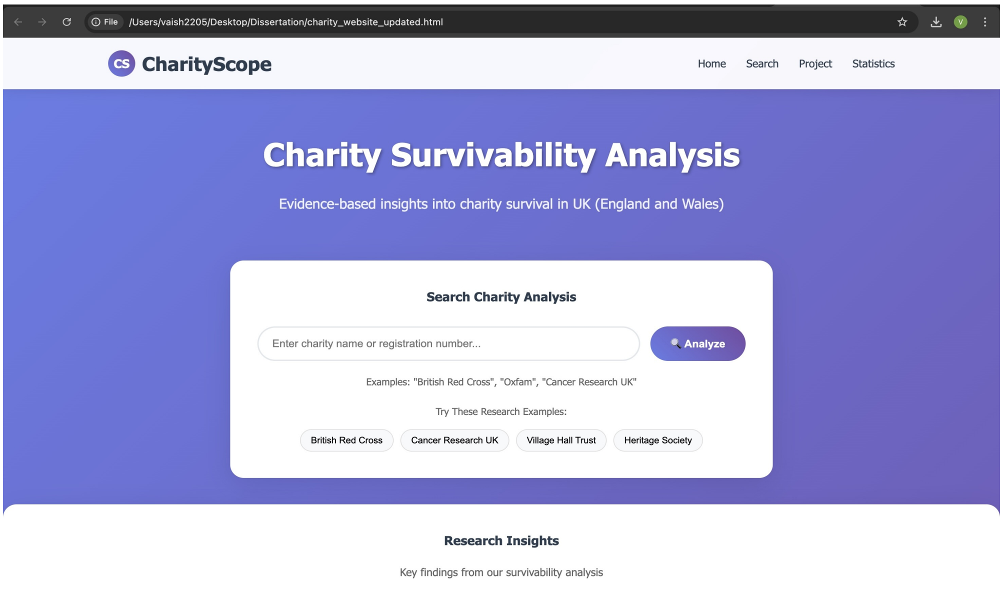
This proof-of-concept demonstrates how the validated logistic regression model can be operationalized for evidence-based decision-making by charity managers, funders, and regulators.

C. Web Application Prototype

These screenshots showcase a prototype web application demonstrating the implementation potential of the charity survival prediction framework. The application translates research findings into accessible, user-friendly tools for charity sector stakeholders.

Application Features:

- Organization search and assessment interface with survival probability calculations
- Risk factor analysis pages covering the five key survival determinants
- Interactive risk assessment tools using the validated logistic regression model
- Educational content explaining methodology and findings for practitioners



Charity Survivability Analysis

Evidence-based insights into charity survival in UK (England and Wales)

Search Charity Analysis

Enter charity name or registration number...

Examples: "British Red Cross", "Oxfam", "Cancer Research UK"

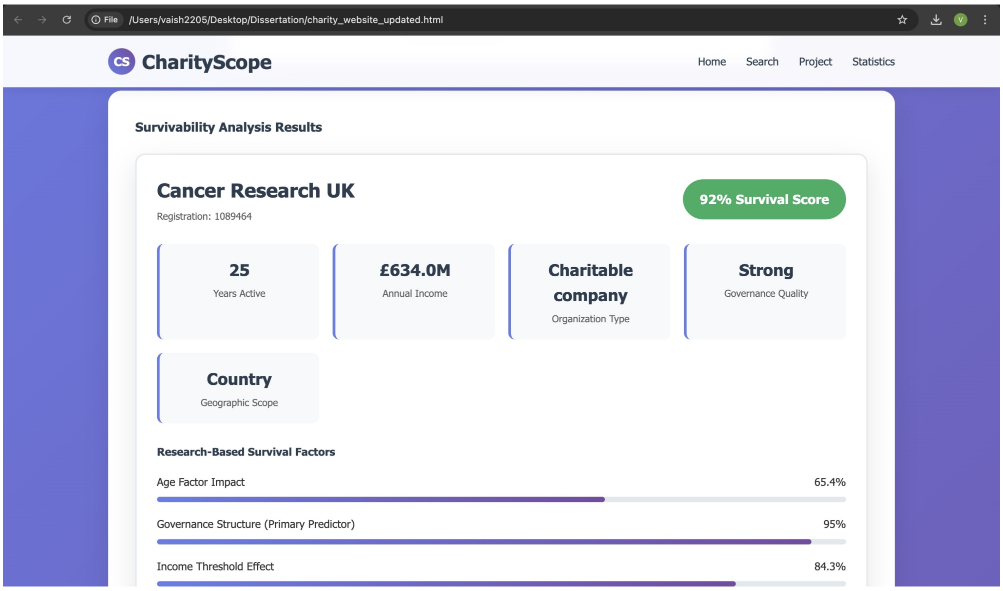
Try These Research Examples:

Research Insights

Key findings from our survivability analysis

Figure 7.2: Appendix: Web Application Interface - Organization Assessment

This prototype illustrates the pathway from academic research to operational early warning systems, demonstrating how key findings (including the extraordinary 99,183% governance quality effects) can be transformed into actionable intelligence for charity sector management and policy development.


File
/Users/vaish2205/Desktop/Dissertation/charity_website_updated.html
CS CharityScope
Home
Search
Project
Statistics

Survivability Analysis Results

Cancer Research UK

Registration: 1089464

92% Survival Score

25	£634.0M	Charitable company	Strong
Years Active	Annual Income	Organization Type	Governance Quality

Country
Geographic Scope

Research-Based Survival Factors

Age Factor Impact 65.4%

Governance Structure (Primary Predictor) 95%

Income Threshold Effect 84.3%

Figure 7.3: Appendix: Web Application Interface - Risk Analysis Dashboard