Predicting Income for Donors and Using Data to Further Understand our Donors

2022 – 2023

MSc Data Science Dissertation

By

Abdulazeez Momoh

@00686172

School of Science, Engineering and Environment

University of Salford

Supervisor: Dr Ian Drumm

September 2023

Abstract

The expansion of data sources and improvements in machine learning techniques offer unprecedented opportunities for charitable organizations to optimize their operations and maximize impact. This project aimed to assist "Make A Wish UK" in effective budget planning and donor segmentation by developing predictive analytics models. Various machine learning models, both traditional and deep learning, were implemented and evaluated to predict donation amounts and classify donors into different segments.

Among the models, Random Forest and XGBoost emerged as the most accurate for predicting donation amounts, while Random Forest Classifier showed the highest accuracy for donor segmentation. These models were not only evaluated based on their technical merits but also on how well they aligned with the project's objectives. The study indicates that adopting such predictive models can significantly improve the charity's financial planning, thereby enhancing its ability to fulfill its mission.

To make the insights accessible, an interactive dashboard was developed using Power BI, featuring real-time data visualizations across various dimensions such as geography, donation categories, and time. This comprehensive approach ensures that the charity can engage in targeted marketing and donor relationship management more effectively.

The project also delves into the ethical considerations and data privacy issues inherent in using predictive analytics in a charitable setting. In summary, this report demonstrates that machine learning can be a potent tool for charities, offering them a data-driven approach for optimized decision-making, but also emphasizes the need for responsible and ethical use of this technology.

Acknowledgements

All praises and thanks be to Allah, the Most Beneficent and the Most Merciful, for His blessings and guidance in my life.

My appreciation goes to my dissertation supervisor, Dr Ian Drumm, for his invaluable guidance and support on this project.

Special thanks to the team at "Make A Wish UK" for granting me access to their data and for their continued cooperation and enthusiasm about the project. Your mission is inspiring, and I am honored to have had the opportunity to contribute in a small way.

To my classmates and friends, your camaraderie and support have made this journey much more enjoyable. Thank you for the insightful discussions, feedback, and the shared moments of triumph and struggle.

Finally, all thanks to my family for their unconditional love, support, and sacrifices. Thank you.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CRISP-DM	Cross-Industry Standard Process for Data Mining
LSTM	Long-Short Term Memory
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
SEMMA	Sample, Explore, Modify, Model, and Assess
KDD	Knowledge Discovery in Databases
ANN	Artificial Neural Network
EDA	Exploratory Data Analysis

Chapter 1: Introduction

1.1 General Overview

An organization's ability to carry out its objective in the philanthropic sector is directly impacted by how its resources are allocated. These funds are frequently obtained through the kindness of a broad donor base, whose donations can differ greatly in terms of frequency and sum given. Strategic planning is therefore necessary for nonprofits like Make A Wish UK (MAW UK), which heavily rely on these donations, to guarantee the best possible use of these resources.

One such organization is Make A Wish UK, which aims to enhance its fundraising efforts by maximizing the value of its existing data. Over the years, Make A Wish UK has transformed many lives by granting wishes to kids and teenagers fighting life-threatening illnesses. Since they are a charity, they heavily rely on donations, so effective fundraising is essential to achieving their goals.

However, predicting contributors' behavior is a challenging task. It entails evaluating variables including the donors' motive, their financial capacity, and the outside forces impacting their contribution decision. Although there isn't a single solution for this complex problem, it does require a strategic strategy to be dealt with.

Non-profits now have the tools to better analyze the behavior of their contributors because to the emergence of data science methodologies like machine learning and predictive analytics.

The application of these cutting-edge methods is not without difficulties, though. It also necessitates having a solid grasp of these approaches' technical components. It necessitates in-depth comprehension of both the technological and organizational needs, as well as the particular context of these approaches. Furthermore, ethical and privacy concerns related to the handling of donor data must be meticulously addressed.

This dissertation explores the use of data-driven methodologies to the problem of forecasting future donations from contributors to MAW UK. The objective is to create a machine learning model that will assist MAW UK in better understanding their donor base, foresee potential future contributions, and as a result, make more smart resource allocation decisions.

This project's possible effects may extend beyond MAW UK. The techniques created and the understandings discovered might be useful to a wide range of charitable organizations, thereby contributing to a broader discourse on resource allocation and donor management within the non-profit sector.

1.2 Problem Statement

Predicting donation patterns and future contributions is a complex issue. Numerous internal and external influences on donors might cause irregularity and unpredictability in the act of giving. Without a data-driven, strategic approach to this problem, outreach attempts may be ineffective, resources may be misallocated, and relationships with potential high-value donors may not be developed. Like many other nonprofits, MAW UK is in need of a solution that would enable them to better understand their donors and anticipate potential future donations. As a result, the issue statement is as follows: *How can MAW UK use data science and machine learning approaches to anticipate future behavior of donors, enabling more strategic resource allocation and improved donor engagement?*

1.3 Aims and Objectives

The primary objective of this life project is to create a data-driven model capable of predicting the potential lifetime contribution of new donors to Make A Wish UK, thereby aiding in the charity's financial planning and donor relationship management. The model will be integrated with a Power BI dashboard, providing an interactive platform for visualizing key insights and predictions.

Here are the specific objectives:

- To conduct exploratory data analysis (EDA) to identify patterns and trends among donors.
- To develop predictive models for forecasting future donor behavior and income.
- To segment donors into different categories based on their donation history.
- To create an interactive dashboard to present the findings and insights.

1.4 Scope of the Study

The study focuses on the donor data from MAW UK, encompassing donor information from 2016 to the present. The scope includes cleaning and preprocessing this data,

performing an EDA, building a predictive model using machine learning, and integrating this model with a Power BI dashboard. Even though MAW UK is the focus of this study, other non-profit organizations can benefit from the techniques and conclusions reached.

1.5 Ethical Considerations

The project involves handling sensitive personal data from donors. Adherence to data protection laws and ethical standards is essential to ensure donor confidentiality. Strict security steps will be implemented to prevent unauthorized access to all data utilized in this study after it has been anonymized. The project will also guarantee openness and transparency, making it apparent to all stakeholders how data will be used, processed, and stored.

1.6 Adopted Approach

Planning and managing each stage of the project development process is essential for success. Data mining methodologies combine the best practices to use when working on a project to provide the best result, and they each offer organized ways to execute job.

The Cross-Industry Standard Process for Data Mining (CRISP-DM) approach will be used for the project.

This widely used methodology gives data mining a structured approach (Chapman et al., 2001). It consists of the following six phase; Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment

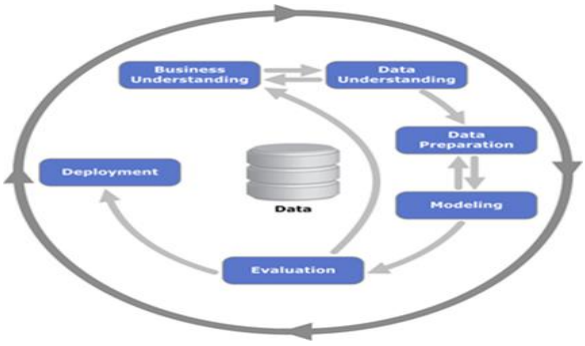

Figure 1. 1: Cross-Industry Standard Process for Data Mining (Smart vision 2020).

The choice of this methodology ensures a systematic approach to the project, from understanding the business needs to deploying the final predictive model. Essentially the adopted approach for this project is fundamentally a data science pipeline composed of several key stages, each contributing to the eventual goal of developing an accurate predictive model and an informative, user-friendly dashboard.

This approach was chosen for its alignment with common data science practices and its adaptability to the specific nature and requirements of the Make A Wish UK project. It ensures a thorough analysis of the data and the creation of a reliable, robust, and useful tool for Make A Wish UK.

1.8 Structure of the Dissertation

This dissertation is organized into five key chapters in order to provide a thorough narrative of the research process and its findings:

- Chapter 1 introduces the research by outlining the problem statement, goals and objectives, scope, methodology, and ethical considerations.
- Chapter 2 provides a complete analysis of the various published papers linked to the application of machine learning in forecasting donor behavior, as well as the methodology used for the research, outlining the CRISP-DM process and the logic behind the chosen machine learning algorithms.
- Chapter 3 provides an in-depth description of the source of the data, structure and description, its limitations, and any biases it may contain. The chapter will also go through exploratory data analysis and pretreatment processes to get the data ready for modeling.
- Chapter 4 This focuses on the analysis, justification of methods used and results also focusing on using different metrics to evaluate the models generated and the overall results of the analysis and the integration with the Power BI dashboard.
- Chapter 5 offers a critical assessment of the project and its findings, as well as the study's limits, potential future improvements, and the research's ramifications for MAW UK and the non-profit sector at large, as well as legal, social, ethical, and professional issues.

Chapter 2: Literature Review and Methodology

The literature review aims to establish a comprehensive understanding of the existing body of knowledge related to the project's focus: predictive analytics in the context of charitable fundraising. This review is vital for situating the current project within the broader discourse, understanding the methodologies, tools, and strategies previously employed, and identifying gaps that the present study might address.

2.1 Introduction

The adoption of data analytics in the nonprofit sector is a relatively recent phenomenon, with growing interest in predictive models for donor behavior (Bennett, 2006). This chapter will conduct a literature analysis on the use of predictive analytics, data mining techniques, and machine learning algorithms in donor prediction and fundraising. It will explore classic predictive models, the growing influence of artificial neural networks, and the need of data visualization in communicating these complicated predictions in an easily digestible way.

While the value of analytics in fundraising is widely recognized, the literature demonstrates a gap in their methodical use and understanding, particularly in the context of UK charities. This review aims to bring various threads of knowledge together and determine how they might be consolidated and implemented in the unique context of Make A Wish UK.

Furthermore, the literature evaluation will contribute to the study's methodological rigor by offering a broader perspective on the strengths and limits of various techniques. It will be used to create and construct the predictive model and data visualization tool in following chapters.

In the following sections, we will delve into the project's various areas of interest. We will start by looking at the role of predictive analytics in fundraising, followed by a discussion on data mining and the CRISP-DM methodology. We will then explore machine learning techniques with a focus on traditional algorithms and artificial neural networks. Finally, we will review the role and significance of data visualization in fundraising analytics.

2.2 Non-profit Fundraising and Donor Behavior

Non-profit organizations (NPOs) are essential to society because they offer assistance and services in sectors like environmental protection, health care, and social services. The ability

to raise money is essential to their operations since it enables them to fulfill their objective. It is crucial to comprehend the variables that affect donor behavior and the efficiency of various fundraising techniques.

The perceived impact of a donor's giving is one of the major elements affecting that behavior. When donors think their contributions will have a big impact, they are more willing to give. This is often communicated through stories and testimonials that highlight the impact of the organizations work (Bekkers & Wiepking, 2011). Therefore, effective storytelling and communication of impact can be a powerful tool in fundraising.

Another important factor is the relationship between the donor and the organization. Donors who feel a strong connection to the organization and its mission are more likely to give and to give more generously. This connection can be fostered through regular communication, opportunities for involvement, and recognition of the donor's contribution (Sargeant & Woodliffe, 2007). Therefore, building and maintaining strong relationships with donors should be a key focus of fundraising efforts.

The fundraising strategy adopted by the organization can also significantly influence donor behavior. Traditional tactics like direct mail, phone calls, and fundraising events are still effective. However, the advancement of digital technology has created new opportunities for fundraising, such as online donating, social media campaigns, and crowd funding. These digital approaches can reach a larger audience and make it easier for consumers to donate. However, they also require a different approach to communication and relationship building (Saxton & Wang, 2014).

Research has also highlighted the potential negative effects of certain fundraising strategies. For example, Damgaard and Gravert (2017) found that while reminders increased donations, they also led to a significant increase in unsubscriptions from the mailing list. This study used a large-scale field experiment with a charity to find that reminders increase donations, but they also substantially increase unsubscriptions from the mailing list. The authors developed a dynamic model of donation and unsubscription behavior with limited attention, which was tested in reduced-form using a second field experiment. They found that reminders are diminishing for the potential donors as non-givers incur a welfare loss of $2.35 for every reminder. The net benefit of every reminder to the charity is $0.18. This suggests that while

reminders can be effective in the short term, they may harm the organization's relationship with donors in the long term.

The effectiveness of fundraising strategies can also be influenced by the broader social, economic, and political context. For example, López de los Mozos, Duarte, and Ruiz (2016) found that changes in the diversification of revenues can negatively impact fundraising efficiency. This study explored how fundraising efficiency is affected by changes in diversification of revenues in non-profit organizations. They found a negative impact on fundraising efficiency when NPOs alter their locus of dependence and change their pattern of diversification. This effect is impacted by organizational size and industry.

Similarly, Zheng, Ni, and Crilly (2018) found that the political ties of charities can influence their success in raising funds from corporate donors. This study developed propositions about how the political ties of charities influence their success in raising funds from corporate donors. They found that organizational political ties, established through formal affiliation with the government, aid fundraising from corporate donors, whereas personal political ties, formed through personal political services of senior leaders of charities, have no such effect.

The fundraising strategy adopted by the organization can also significantly influence donor behavior. Traditional methods such as direct mail, telephone calls, and fundraising events continue to be effective. However, the rise of digital technology has opened up new avenues for fundraising, including online giving, social media campaigns, and crowdfunding. These digital methods can reach a wider audience and provide more convenient ways for people to donate. However, they also require a different approach to communication and relationship building (Saxton & Wang, 2014).

One particular challenge is donor retention, defined as the ability of a non-profit organization to keep a donor over time. According to a report by the Fundraising Effectiveness Project (2018), the average donor retention rate in the U.S. is below 50%, indicating that over half of all donors do not donate again to the same organization within one year.

Improving donor retention is a strategic priority for non-profit organizations. Retained donors typically contribute more over their lifetime and are more likely to promote the organization through word-of-mouth, enhancing the organization's reputation and reach (Sargeant & Jay, 2014). Moreover, it is generally more cost-effective to retain existing donors than to acquire new ones, making donor retention critical for non-profit sustainability.

The quest for donor retention has sparked interest in understanding donor behavior. Non-profits are increasingly leveraging data to gain insights into donors' motivations, giving patterns, and preferences. These insights enable non-profits to tailor their communication strategies, cultivate meaningful relationships with donors, and ultimately enhance donor loyalty and retention.

Successful fundraising requires a deep understanding of donor behavior and the factors that influence it. It also requires a strategic approach to fundraising that considers the potential benefits and drawbacks of different strategies and adapts to the changing context. As we move forward with our project, these insights will guide our approach to predicting future donations and developing effective fundraising strategies for Make A Wish UK.

2.3 Role of Predictive Analytics in Fundraising

In the increasingly data-intensive landscape of the 21st century, predictive analytics has emerged as an invaluable tool for various applications. Non-profit organizations are turning to predictive analytics to streamline their fundraising efforts, leveraging data to forecast future trends, behaviors, and events. This innovative approach is showing impressive results in identifying potential donors and estimating their propensity and capacity to donate.

Predictive analytics' utility in fundraising was convincingly demonstrated in a study conducted by Wang and Shih (2009). They employed predictive models to determine the likelihood of donation, potential donation amounts, and the probability of individuals becoming regular donors. Their results showed that predictive analytics significantly increased the effectiveness of fundraising campaigns by helping organizations target their efforts more strategically.

Similarly, Buelens et al. (2012) conducted a study demonstrating how charities could use predictive models to bolster donor retention strategies. Their findings illustrated that data-driven insights could enhance understanding of donor behavior, leading to better communication and improved retention.

More recently, the University of San Francisco used predictive analytics to boost their annual fundraising campaign. Using their alumni database, the university built predictive models to identify the most likely donors, forecast donation amounts, and improve their outreach strategies. The campaign resulted in a significant increase in donations compared to previous years, underscoring the potential of predictive analytics in enhancing fundraising effectiveness.

Non-profit organization *DonorsChoose.org,* a platform where public school teachers from across the U.S. post classroom project requests for donations, also successfully utilized predictive analytics. They built models to predict the probability of a donor returning to the platform and donating again. By adjusting their engagement strategies based on these insights, they were able to significantly improve their donor retention rate.

However, it's important to note that predictive analytics not only increases donation amounts but also allows charities to foster a more personalized and meaningful relationship with their donors. By understanding and anticipating donor behavior, non-profit organizations can build stronger connections with their donors, enhancing their engagement and loyalty over time (Bryant, 2013).

While predictive analytics offers significant benefits, it's critical to acknowledge its limitations. The accuracy of the predictions heavily depends on the quality and relevance of the underlying data. Plus, ethical considerations surrounding data privacy and usage cannot be overlooked and should always be addressed with sensitivity and caution.

Predictive analytics is critical in modern fundraising. It not only improves fundraising efficiency but also deepens donor engagement by helping non-profit organizations to efficiently understand, anticipate, and respond to donor behavior.

2.4 Data Mining and Machine Learning in Fundraising

Methodological Approach

The methodological approach for this study is centered on data mining and machine learning techniques. The goal of these approaches is to extract actionable insights from the acquired data and then utilize those insights to forecast total lifetime donations from new donors. These approaches were chosen based on their shown efficacy in previous projects as well as their suitability for the type of the dataset at hand.

Data mining is an interdisciplinary subfield of computer science and statistics with the overarching objective of extracting information from a dataset and transforming it into a usable structure. It is the discovery of patterns, correlations, and trends in massive datasets.

Data mining techniques will be utilized in this project to reveal links and patterns in donor behavior, which are critical for successful decision-making and strategic planning in fundraising.

Statistical models and algorithms are used by computers to carry out tasks without explicit instructions, relying instead on patterns and inference. Machine learning is a subset of artificial intelligence.

In choosing the specific machine learning techniques, careful consideration was given to the nature of the task. As the task is to predict a continuous value (the total lifetime donations), regression techniques are appropriate. Multiple linear regression, gradient boost regression, and random forest regression are some of the techniques that will be considered as well as deep learning models. Also to segment the donors into different categories, classification techniques are appropriate. Logistic regression classifier, Random forest classifier, gradient boost and deep learning models were considered. These techniques were chosen due to their ability to handle a large number of variables and their applicability in predicting numerical outcomes.

The overall workflow of this project involves a series of steps: data cleaning, exploratory data analysis, feature engineering, model building, model evaluation, and finally, dashboard creation. Data cleaning ensures that the dataset is free of errors and inconsistencies that could affect the model's performance. Exploratory data analysis involves probing the dataset to understand its characteristics and uncover initial insights. Feature engineering is the process of creating new features that can enhance the model's ability to learn. Model building involves training a machine learning model on the dataset, while model evaluation checks the performance of the model using suitable metrics.

Finally, an interactive dashboard will be built using Power BI to visualize the findings and insights from the analysis.

In summary, the methodological approach for this study involves a comprehensive combination of data mining and machine learning techniques to extract meaningful insights from the data and create a predictive model for future donations. Each stage of this approach has been thoughtfully planned to ensure that the final results are robust, reliable, and can aid in effective decision-making.

2.4.1 Data Mining and Machine Learning

Data mining and machine learning are two interconnected fields that have revolutionized the way we analyze and interpret data.

- **Data Mining**

Data mining is the process of discovering patterns and knowledge from large amounts of data. The term is a misnomer, for it implies mining from data, not mining for data (Han, Pei, & Kamber, 2011). It involves the use of sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large pre-existing databases.

Data mining techniques can be broadly classified into two types: descriptive and predictive. Descriptive data mining techniques, such as clustering and association rules, are used to explore the data and find patterns that can help understand the data better. Predictive data mining techniques, such as classification and regression, are used to build models that can predict future outcomes based on historical data.

- **Machine Learning**

Machine learning, a subset of artificial intelligence, involves the development of algorithms that allow computers to learn from and make decisions or predictions based on data (Bishop, 2006).

Machine learning can be categorized into three types: supervised learning, unsupervised learning, and reinforcement learning.

- **Supervised Learning**

In supervised learning, the model is trained on a labeled dataset. This means that the dataset includes both the input data and the correct output. The model learns to map the input to the output during the training process. Once the model is trained, it can be used to predict the output for new, unseen input data. Common supervised learning algorithms include linear regression for regression problems and logistic regression, decision trees, and support vector machines for classification problems.

- **Unsupervised Learning**

In unsupervised learning, the model identifies patterns in an unlabeled dataset. This means that the dataset includes the input data but not the output. The model must discover the underlying structure in the data without any guidance. Common unsupervised learning algorithms include k-means for clustering problems and principal component analysis for dimensionality reduction problems.

- **Reinforcement Learning**

In reinforcement learning, the model learns by interacting with its environment. The model, also known as an agent, takes actions in an environment to achieve a goal. The agent receives feedback from the environment in the form of rewards or penalties. The agent's goal is to learn a policy, which is a strategy for choosing actions that maximize the total reward over time (Sutton & Barto, 2018).

The particular issue at hand, the type of data, and the desired result all influence the decision of the data mining methods and machine learning algorithms to use.

2.4.2 Data Mining and Machine Learning Techniques in Fundraising

The application of data mining and machine learning techniques in predicting income from donors has been a subject of interest in recent years. Data mining and machine learning techniques provide potent tools for fundraising organizations, enabling them to analyze large datasets and extract meaningful patterns and predictions. These techniques can help predict donor behavior and tailor strategies to maximize fundraising efficiency and effectiveness.

- **Decision Trees and Regression Models**

Decision trees and regression models are two common techniques used for their ability to create understandable and interpretable models. A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and leaf nodes represent decisions.

In the context of fundraising, a decision tree might be used to segment donors based on their past giving behavior, demographic characteristics, and other relevant attributes. For example, a tree might branch based on whether a donor has given more than a certain amount in the past, with further branches based on age, income level, or geographic location.

Regression models, on the other hand, can provide quantitative estimates of the relationship between various factors and the amount a donor is likely to give. For instance, a regression model might indicate that for each additional year a donor has been associated with an organization, their expected donation amount increases by a certain percentage.

Larose and D. T. Larose (2014) applied a combination of regression models and decision trees in their study on fundraising. They employed these models to analyze the demographics and past giving behavior of donors, improving the accuracy of predicting donation amounts.

- **Clustering**

Clustering algorithms segment donors into distinct groups, allowing for targeted fundraising strategies. These algorithms group donors based on their similarities in certain defined categories like giving behavior, demographic characteristics, or engagement with the organization.

- **Artificial Neural Networks(ANN)**

Artificial neural networks, inspired by biological neural networks, can model complex, non-linear relationships. They consist of interconnected layers of nodes or "neurons" that can process and transmit information, making them powerful tools for prediction.

In a study by Coussement and Van den Poel (2008), they applied artificial neural networks to predict donor churn, an essential aspect of donor retention. They used a variety of factors, including the donor's giving history, interaction with the organization, and demographic information, to train the network. The model was able to successfully predict whether a donor would stop donating in the future, allowing the organization to take proactive steps to retain those donors.

Apart from these, several studies and projects have demonstrated the effectiveness of data mining and machine learning in fundraising. For instance, a study by Zhang, Zhu, and Wang (2010) used data mining techniques to analyze the donation data of a university alumni association. The study used clustering algorithms to segment the donors into different groups and found that this segmentation could help the association develop more targeted fundraising strategies.

In another example, the non-profit organization DonorsChoose.org used machine learning algorithms to predict which projects are most likely to get fully funded. The organization used a variety of features, including the amount of the request, the poverty level of the school, and the subject of the project, to train their model. The model was able to accurately predict the

likelihood of a project getting fully funded, helping the organization prioritize its resources and efforts.

Bhattacharya and Bandyopadhyay (2011) utilized an ensemble of decision tree, logistic regression, and artificial neural networks to predict the lifetime value of a donor. By considering factors like donation frequency, average donation amount, and time since the last donation, their model could help organizations identify high-value donors and devise strategies to engage and retain them.

Similarly, Tempel and de la Torre (2016) used machine learning techniques, including decision trees and clustering, to segment donors for a major museum. By segmenting donors based on their giving history, the museum was able to create tailored fundraising strategies for each group, improving the effectiveness of their fundraising efforts.

Museli (2012) discusses the use of machine learning methods in predicting behavior in different cases based on previously collected observations. Although the context is biomedical data, the principles of using machine learning for prediction can be applied to predicting donor behavior. Machine learning algorithms can be trained on historical donation data, learning patterns and relationships between different variables. This learned knowledge can then be used to predict future donations, allowing organizations to better plan their fundraising strategies.

In a cross-country study on donor responses to fundraising appeals for the COVID-19 fight, Bin-Nashwan et al. (2022) explore the motivations driving donors to engage in fundraising appeals launched through social networking sites. Understanding these motivations can provide valuable insights into donor behavior, which can be used to improve the accuracy of predictive models. For instance, if it is found that donors are more likely to donate to causes that align with their personal values, this information can be used to tailor fundraising appeals to individual donors, potentially increasing donation amounts.

Coulter, Baingana, and Mukakamari (2019) discuss the use of machine learning algorithms to predict donor status. They used a dataset from an institution of higher learning, training different machine learning models on the data and comparing their performance. The models were able to predict donor status with a high degree of accuracy, demonstrating the potential of machine learning in this context. This study provides a practical example of how machine learning can be used in the context of fundraising, showing that these techniques can be successfully applied to real-world data.

Mittal (2021) reviews different supervised machine learning algorithms for classifying donors for charity. The paper provides insights into the specific algorithms that can be used for predicting donor behavior, including decision trees, random forests, and support vector machines. Each of these algorithms has its strengths and weaknesses, and the choice of algorithm can depend on various factors, such as the size and nature of the dataset, the complexity of the relationships between variables, and the specific objectives of the prediction task.

Despite the promising results of these studies, there are challenges and limitations to the use of data mining and machine learning in predicting income from donors. These include the quality and availability of data, the complexity of donor behavior, and the risk of overfitting, among others. However, with careful data preprocessing, feature selection, and model validation, these challenges can be mitigated.

Looking forward, there are many potential directions for future research in this area. These include the development of more sophisticated machine learning models, the integration of additional types of data (such as social media data), and the exploration of other aspects of donor behavior. As technology continues to advance and more data becomes available, the potential for data mining and machine learning in predicting income from donors is likely to continue to grow.

As you can tell, data mining and machine learning techniques have shown significant promise in improving fundraising effectiveness. By providing valuable insights into donor behavior, these techniques allow organizations to better understand their donors and devise more effective fundraising strategies.

2.4.3 CRISP-DM in Predictive Analytics for Fundraising

Cross-Industry Standard Process for Data Mining (CRISP-DM) is a robust, structured approach to data mining that is particularly effective for predictive analytics projects. CRISP-DM is a cyclical process comprising six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

In the context of predictive analytics for fundraising, the CRISP-DM methodology would be used as follows:

- **Business Understanding:** This phase involves understanding the goals and requirements of the fundraising project from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan.
- **Data Understanding:** The data understanding phase starts with data collection, followed by activities to get familiar with the data, identify data quality issues, discover initial insights, and detect interesting subsets to form hypotheses for hidden information.
- **Data Preparation:** This phase includes all the activities necessary to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include cleaning, transforming, and selecting data.
- **Modeling**: In this phase, various modeling techniques are selected and applied. Typically, there are several techniques that might be applicable, and it's necessary to select the best one based on the nature of the data and the goal of the project.
- **Evaluation:** Once one or more models are built that appear to have high-quality predictions based on their ability to find patterns within the data, it is crucial to thoroughly evaluate them before deploying to ensure they meet the project's objectives.
- **Deployment**: The knowledge gained from the model is organized and presented in a way that the customer can use it. It may be as simple as a report, or it may be a more dynamic data-driven predictive analytics tool.

Applying the CRISP-DM methodology to a predictive analytics project for fundraising ensures a systematic, structured approach, maximizing the chances of achieving reliable, useful results.

2.5 Data Visualization in Fundraising

2.5.1 Data Visualization

Data visualization is the practice of representing data in a graphical or pictorial format. It allows for easier understanding of complex data, facilitates the detection of patterns, trends, and outliers, and aids in the formulation of insights (Yau, 2013). It encompasses a set of techniques used to encode numbers and data in graphs and charts, enabling a more profound comprehension of the information presented. Its importance lies in its ability to communicate a narrative uniquely and effectively, making it an indispensable tool in data-driven decision-

making processes. Data visualization is particularly essential in the modern era of big data, where the sheer volume and complexity of information can be overwhelming.

There are several common methods and techniques in data visualization. These include, but are not limited to, bar charts, pie charts, line graphs, scatter plots, heat maps, and more complex forms such as treemaps and network diagrams. The choice of visualization method depends on the nature of the data and the specific insights one seeks to derive.

From basic pie and bar charts to sophisticated interactive maps and complex network diagrams, the field of data visualization encompasses a vast range of techniques. These techniques are specifically designed to exploit human visual perception abilities, aiding in identifying patterns, correlations, and trends that might go unnoticed in text-based data (Kirk, 2016).

However, the power of data visualization comes with responsibility. Effective visualizations must maintain a balance between aesthetic appeal and the faithful representation of data. Misleading or inaccurate visualizations can distort the data's underlying message, leading to misinterpretations and erroneous conclusions (Cairo, 2013).

Data visualization is not without its ethical considerations. Accuracy and honesty are paramount. Misrepresentation of data, whether intentional or not, can lead to incorrect interpretations and misinformed decisions (Resnik, 2015). Moreover, data privacy is another significant concern, especially when visualizing sensitive information.

Data visualization also plays a pivotal role in communicating results to stakeholders, especially those without a technical background. By presenting data visually, complex information can be communicated clearly, making it easier for decision-makers to understand and act on this information.

2.5.2 Data Visualization in Fundraising

In the domain of fundraising, data visualization plays a crucial role in conveying the impact and value of donors' contributions, planning and strategizing fundraising campaigns, and making data-driven decisions. From internal decision-making to external communication, visualization is at the heart of understanding and effectively leveraging fundraising data. Hence, it can be harnessed to aid in strategic planning, monitor progress, and communicate with a wide array of stakeholders including donors, employees, and the public (Knaflic, 2015).

Fundraising teams often handle vast volumes of data, including donor demographics, donation amounts, donation frequency, and campaign outcomes. Visualizing this data allows for clearer comprehension and analysis, making it easier to identify trends and patterns, such as common characteristics of high-value donors, or times of the year when donations surge.

There have been many successful uses of data visualization in fundraising.

An example is the Global Giving platform, which uses data visualization to track the progress of projects they fund. The platform uses a variety of charts and graphs to represent the collected data, including funds raised, remaining goals, and geographic location of donors (GlobalGiving, 2023).

Moreover, nonprofits like Charity: Water use data visualization to show donors exactly where their money is going and the impact it's making. They offer visuals showing the number of water projects funded, people served, and countries worked in. This kind of transparent data visualization can increase trust and boost donor retention (Charity: Water, 2023).

However, with the power of data visualization comes the responsibility of using it ethically. As Rogowitz and Treinish (1998) caution, visualizations can be manipulated to misrepresent data, leading to misinformation. Thus, it is crucial to represent data accurately and ethically.

Additionally, interactive data visualization tools can provide a more dynamic and engaging way for stakeholders to explore fundraising data. For instance, The Gates Foundation uses interactive visualizations to communicate their work's impact, thereby encouraging further support (Gates Foundation, 2022).

Data visualization is an invaluable tool in fundraising. It not only enables non-profits to understand their data and make informed decisions but also plays a key role in building transparency and trust with donors.

2.6 Ethical Considerations in Data Analysis for Fundraising

Fundraising analytics, like any other application of data analysis, is not exempt from ethical considerations. When dealing with personal donor information, non-profit organizations must be vigilant in adhering to ethical and legal norms regarding privacy, consent, and the use of personal data.

Privacy is a major concern in fundraising analytics. Non-profit organizations have access to sensitive information about donors, including contact details, financial transactions, and sometimes, personal interests and affiliations. This information, while crucial for predictive analytics, is also sensitive and personal. Breaches in data security could have serious consequences, damaging the organization's reputation and potentially causing harm to the donors involved.

Informed consent is another key ethical principle. This means that donors should know and understand how their data will be used, including any predictive analytics or data mining applications. They should also have the opportunity to opt out if they do not wish their data to be used in this way.

Finally, non-profits must ensure that their use of fundraising analytics does not lead to unfair or discriminatory practices. For example, data mining algorithms can sometimes replicate and amplify existing biases in the data, leading to unfair outcomes.

Organizations must, therefore, approach fundraising analytics with care, ensuring that they strike the right balance between maximizing their fundraising potential and respecting the rights and interests of their donors.

2.7 Summary and Identified Gaps

In summary, the literature review has illuminated a significant body of research on predictive analytics, data mining, and data visualization in the fundraising sector. Scholars and practitioners have demonstrated the efficacy of predictive analytics and data mining in enhancing fundraising efforts, notably through their application to donor retention and lifetime value prediction.

The key findings from the reviewed literature include:

- Predictive analytics and data mining techniques can help nonprofit organizations better understand donor behavior, predict future donations, and enhance donor retention.
- Data visualization aids in the presentation and understanding of complex data patterns and insights derived from predictive analytics and data mining. It enables the communication of these insights to diverse stakeholders.

- Fundraising strategies are shifting towards a data-driven approach, and the integration of predictive analytics, data mining, and data visualization has proven beneficial in augmenting these strategies.

Despite these advancements, identified gaps persist in the current body of knowledge:

There is limited empirical research exploring the integration of predictive analytics, data mining, and data visualization in fundraising strategies within nonprofit organizations. More specifically, studies that demonstrate how these techniques are used in concert to optimize fundraising outcomes are relatively scarce.

A lack of context-specific studies is also apparent. Different types of nonprofit organizations face unique challenges in fundraising, and research into how predictive analytics, data mining, and data visualization are used in these specific contexts is lacking.

Lastly, there is a need for more longitudinal studies to understand the long-term impacts and benefits of integrating predictive analytics, data mining, and data visualization in fundraising strategies.

The present project aims to fill these gaps by investigating the integration of predictive analytics, data mining, and data visualization in fundraising efforts within nonprofit organizations. More specifically, it will conduct a context-specific study to shed light on the application of these techniques in a selected nonprofit organization.

Chapter 3: Data Understanding, Preparation and EDA

This chapter aims to provide a comprehensive understanding of the dataset used in this project. It will cover the source of the data, structure and description, its limitations, and any biases it may contain. The chapter will also delve into the exploratory data analysis and preprocessing steps taken to prepare the data for modeling.

3.1 Data Understanding

3.1.1 Data Source

The Dataset was provided by “Make A Wish Uk” (a non-profit organization) from their Salesforce. This data contains multiple years' worth of donation records from 2016 to present, which include features such as names, addresses, unique IDs, payment amounts, payment source codes, types of income, contact record types, and gender. Each of these features forms the basic building blocks of our study and understanding their properties, types (e.g., categorical, numerical), and relevance to the project objectives is a crucial initial step. The data was provided in three csv files titled; 2016-2019, 2019-2021 and 2021-Present. A metadata was equally provided.

3.1.2 Data Structure and Description

The combined dataset comprises 34 variables (columns) and 884,982 rows. Each row in this dataset represents a unique donation transaction, which will be used to analyze and predict donor behavior. Table 3.1 below outlines all the attributes in the dataset and their respective descriptions. These attributes serve as the independent variables. A class label or dependent variable, termed "Donor Category," was later generated based on the 'Payment Amount' field. If the payment amount is above a certain threshold, the record is labeled as a "High-value donor"; if below, it is labeled as a "Low-value donor." With this dataset, we aim to solve the classification problem of categorizing donors into different segments and also focus on the individual or collective impact of attributes like 'Payment Source Code,' 'Gender,' 'Recruitment Source.1,' etc., on donor behavior.

Table 3. 1 Attributes, their types and description.

Attribute	Data Type	Description
Payment Number	Numerical	Unique identifier for each payment/donation.
Contact ID	Alphanumerical	Unique identifier for each donor.
Salesforce Number	Alphanumerical	Another unique identifier, for each donor.
Payment Source Code: Source Code	Categorical	Code representing the source of the payment.
Salutation	Categorical (Text)	Title or form of address (e.g., Mr., Mrs., Dr.).
Payment Amount Received Currency	Categorical (Text)	Currency in which the payment was received.
Payment Amount Received	Numeric (Float)	Amount of the payment received.
Payment Amount	Numeric (Float)	Amount of the payment.
Payment Date	Date	Date of the payment.
Total Gifts Last Year Currency	Categorical (Text)	Currency of total donation last year.
Total Gifts Last Year	Numeric (Float)	Total amount of donation last year.
Last Gift Date	Datetime	Date of the last donation
Total Gifts Two Years Ago Currency	Categorical (Text)	Currency of total donation two years ago.
Total Gifts Two Years Ago	Numeric (Float)	Total amount of donation two years ago.
Total Number of Gifts	Numerical	Total number of gifts given by the donor.
Mailing Address	Text	Complete mailing address of the donor.
Mailing Street	Text	Street component of the mailing address.
Mailing City	Text	City component of the mailing address.

Mailing Zip/Postal Code .	Text	component of the mailing address.
Mailing Country (text only)	Text	Country component of the mailing address.
Mailing State/Province	Text	component of the mailing address.
Recruitment Source Categorical	Categorical (Text)	Source through which the donor was recruited.
Primary Campaign Source: Campaign Name	Categorical (Text)	Name of the primary campaign source.
Source Code: Source Code	Object	Another code for the source of the payment.
Primary Campaign Source: Event Code: Event Code	Categorical (Text)	code of the primary campaign source.
Primary Campaign Source: Campaign Name.1	Categorical (Text)	Another name for the primary campaign source.
Primary Campaign Source: Parent Campaign: Campaign Name	Categorical (Text)	Parent campaign of the primary campaign source.
Payment Source Code: Type of Income	Categorical (Text)	Type of income (e.g., Cash, Regular giving, supporter led).
Contact Record Type	Categorical (Text)	Type of contact (e.g., individual, organization).
Opportunity Record Type	Categorical (Text)	Type of opportunity related to the payment.
Gender	Categorical (Text)	Gender of the donor.
Primary Language	Categorical (Text)	Primary language of the donor.
Age	Numerical	Age of the donor.
Recruitment Source.1	Categorical (Text).	The avenue in which the donation was made.

3.2 Data Limitations and Biases

Data privacy and bias are pivotal concerns that demand continuous scrutiny, especially when dealing with machine learning and data science. Biases in a dataset can lead to skewed or inaccurate results, thereby affecting the integrity of the model and its applicability (Aslan, 2021). This section aims to identify and discuss the potential biases and limitations present in the dataset used for this research.

Dataset-Specific Observations

The dataset used in this research primarily consists of donation data, including personal and financial details. While it does not contain overly sensitive information like medical records, it does have several limitations and biases: Data privacy and bias are pivotal concerns that demand continuous scrutiny, especially when dealing with machine learning and data science. Biases in a dataset can lead to skewed or inaccurate results, thereby affecting the integrity of the model and its applicability (Aslan, 2021). This section aims to identify and discuss the potential biases and limitations present in the dataset used for this research.

3.2.1 Limitations

Missing Values

Some features like 'Gender' and 'Mailing Address' have missing values, which could potentially skew the analysis and predictions. These missing values will need to be addressed either through imputation or by removing the records altogether.

Currency Variation

The 'Payment Amount' feature contains amounts in multiple currencies. This adds an extra layer of complexity as the amounts need to be standardized to a common currency for a fair comparison and analysis.

3.2.2 Biases

Gender Bias

The dataset appears to have a gender bias, with more records for one gender over another. This imbalance could lead to biased predictions and insights. For example, if the dataset contains more male donors than female donors, the model might be more accurate in predicting the behavior of male donors.

Socio-Economic Bias

Although not immediately apparent, there may be a socio-economic bias in the dataset. Donors from higher income brackets may be overrepresented, which could skew the analysis and predictions.

By addressing these limitations and biases, we aim to build a robust data pipeline that will serve as the foundation for accurate and unbiased predictive models. The next sections will delve into the exploratory data analysis and preprocessing steps that were taken to mitigate these issues.

3.2.3 Ethical Considerations

Data Privacy: The dataset contains sensitive information like 'Mailing Address', which must be handled carefully to comply with data protection laws.

Fairness: The potential gender and geographical biases in the dataset raise concerns about the fairness of any models trained on this data.

Transparency: The high cardinality and missing values in several columns may affect the transparency of the model's decision-making process.

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a foundational step in the data science pipeline, aimed at understanding the structure and patterns within a dataset. This process employs statistical and visualization techniques to identify anomalies, outliers, and relationships among the variables (Liu et al., 2020).

The objective of this section is to answer key questions about the dataset, which will inform subsequent modeling and analysis.

Research Questions

- What kind of information are we dealing with?
- What data types are contained in the dataset?
- Is the dataset complicated?
- Are any values or entities missing?
- Is there a connection between the dataset's properties or variables?
- Is there a difference in the data attributes' dimensions or scales?
- Which characteristics are most relevant for developing the model?
- Is there anything that may be included in the modeling stage to help with model optimization?

3.3.1 Determining the Data Types and Complexity

The datasets were initially loaded into a Jupyter Notebook and merged for analysis. Utilizing the info function, a summary was generated to identify the data types and complexity of the dataset. The dataset primarily consists of numerical values, with exceptions like 'Salutation' and 'Gender,' which are categorical. Additionally, 'Payment Date' and 'Last Gift Date' are datetime64[ns] types, adding a temporal dimension to the dataset.

While numerical variables like 'Payment Amount Received' and 'Total Gifts Last Year' are continuous, categorical variables like 'Gender' and 'Recruitment Source.1' are nominal. It's worth noting that the dataset contains a mix of object, float64, int64, and datetime64 [ns] types, making it a complex dataset to analyze.

A descriptive statistical analysis was performed to understand the central tendencies and distributions of the variables. Metrics such as mean, standard deviation, and percentiles were calculated for numerical variables. The dataset comprises 35 variables and 884982 observations, making it a large and complex dataset for analysis.

Importing the required Libraries

```
!pip install geopy

import pandas as pd
import numpy as np
import json
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib.ticker as ticker
import requests

from urllib3.util.retry import Retry
from requests.adapters import HTTPAdapter
from geopy.geocoders import Nominatim
from geopy.exc import GeocoderUnavailable, GeocoderTimedOut

from urllib.request import urlopen
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.model_selection import KFold

from keras.models import Sequential
from keras.layers import Dense
```

Figure 3. 1: Importing the required libraries

Loading the Datasets

```
# Loading the 2016-2019 dataset
donor_16_19 = pd.read_excel(r'C:\Users\HP\Documents\Course work\Project\2016-2019.xlsx')

# Loading the 2019-2021 dataset
donor_19_21 = pd.read_excel(r'C:\Users\HP\Documents\Course work\Project\2019-2021.xlsx')

# Loading the 2021-2023 dataset
donor_21_23 = pd.read_excel(r'C:\Users\HP\Documents\Course work\Project\2021-present.xlsx')

# Merging the datasets
merged_donors_df = pd.concat([donor_16_19, donor_19_21, donor_21_23], ignore_index=True)
```

Figure 3. 2: Loading and merging the dataset.

Exploring the Dataset

```
# Viewing the first few rows of the dataset
merged_donors_df.head(5)
```

	Payment Number	Contact ID	Salesforce Number	Payment Source Code: Source Code	Salutation	Payment Amount Received Currency	Payment Amount Received	Payment Amount Currency	Payment Amount	Payment Date	...	Primary Campaign Source: Campaign Name.1	Primary Campaign Source: Parent Campaign: Campaign Name	Payment Source Code: Type of Income	
0	PMT-2027230	0034J00000fAOP5	SF-025146	KNOREG	Mrs	GBP	5.0	GBP	5.0	2018-12-07	...	NaN	NaN	Regular Giving	In
1	PMT-2796235	0034J00000fAOP5	SF-025146	KNOREG	Mrs	GBP	5.0	GBP	5.0	2017-04-05	...	NaN	NaN	Regular Giving	In
2	PMT-3017483	0034J00000fAOP5	SF-025146	KNOREG	Mrs	GBP	5.0	GBP	5.0	2018-03-05	...	NaN	NaN	Regular Giving	In
3	PMT-3017724	0034J00000fAOP5	SF-025146	KNOREG	Mrs	GBP	5.0	GBP	5.0	2016-10-05	...	NaN	NaN	Regular Giving	In
4	PMT-3026573	0034J00000fAOP5	SF-025146	KNOREG	Mrs	GBP	5.0	GBP	5.0	2019-06-07	...	NaN	NaN	Regular Giving	In

5 rows × 36 columns

Figure 3. 3: Retrieving the first few rows in the dataset

This provides a snapshot of the data to understand the structure of your dataset.

```
# Viewing the number of rows and columns in the dataset
merged_donors_df.shape
```

```
(884982, 36)
```

```
# To get a concise summary of the dataframe
merged_donors_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 884982 entries, 0 to 884981
Data columns (total 36 columns):
 #   Column                                                Non-Null Count   Dtype
---  ------                                                --------------   -----
 0   Payment Number                                        884982 non-null  object
 1   Contact ID                                            884982 non-null  object
 2   Salesforce Number                                     884982 non-null  object
 3   Payment Source Code: Source Code                      884953 non-null  object
 4   Salutation                                            660462 non-null  object
 5   Payment Amount Received Currency                      884982 non-null  object
 6   Payment Amount Received                               884982 non-null  float64
 7   Payment Amount Currency                               884981 non-null  object
 8   Payment Amount                                        884981 non-null  float64
 9   Payment Date                                          884982 non-null  datetime64[ns]
 10  Total Gifts Last Year Currency                        884982 non-null  object
 11  Total Gifts Last Year                                 884982 non-null  float64
 12  Last Gift Date                                        834144 non-null  datetime64[ns]
 13  Total Gifts Two Years Ago Currency                    884982 non-null  object
 14  Total Gifts Two Years Ago                             884982 non-null  float64
 15  Total Number of Gifts                                 884982 non-null  int64
 16  Mailing Address                                       825105 non-null  object
 17  Mailing Street                                        823455 non-null  object
 18  Mailing City                                          714645 non-null  object
 19  Mailing Zip/Postal Code                               822552 non-null  object
 20  Mailing Country (text only)                           638576 non-null  object
 21  Mailing State/Province                                516876 non-null  object
 22  Recruitment Source                                    132193 non-null  object
 23  Primary Campaign Source: Campaign Name                82427 non-null   object
 24  Source Code: Source Code                              883687 non-null  object
 25  Primary Campaign Source: Event Code: Event Code       82158 non-null   object
 26  Primary Campaign Source: Campaign Name.1              82427 non-null   object
 27  Primary Campaign Source: Parent Campaign: Campaign Name  77414 non-null  object
 28  Payment Source Code: Type of Income                   884786 non-null  object
 29  Contact Record Type                                   884982 non-null  object
 30  Opportunity Record Type                               884982 non-null  object
 31  Gender                                                671058 non-null  object
 32  Primary Language                                      9630 non-null    object
 33  Age                                                   1513 non-null    float64
 34  Last Gift Amount                                      881325 non-null  float64
 35  Recruitment Source.1                                  291160 non-null  object
dtypes: datetime64[ns](2), float64(6), int64(1), object(27)
memory usage: 243.1+ MB
```

Figure 3. 4: Different data types in the dataset.

It indicates the data types. Knowing the data types of each column is essential for selecting the appropriate analysis methods and visualization techniques.

```
# Descriptive Statistical Analysis: summarize the central tendancy and shape of the dataset
merged_donors_df.describe()
```

	Payment Amount Received	Payment Amount	Total Gifts Last Year	Total Gifts Two Years Ago	Total Number of Gifts	Age	Last Gift Amount
count	884982.000000	884981.000000	884982.000000	884982.000000	884982.000000	1513.000000	881325.000000
mean	216.976333	39.991172	69.101284	213.293849	176.827980	25.399868	39.687378
std	2305.613018	1759.502400	1897.160891	1580.407518	789.980276	17.646647	1216.086383
min	-180.000000	-7753.170000	-3.000000	0.000000	0.000000	1.000000	-2500.000000
25%	3.840000	3.840000	0.000000	0.000000	17.000000	9.000000	2.500000
50%	5.000000	5.000000	0.000000	0.000000	66.000000	23.000000	5.000000
75%	12.500000	10.000000	15.000000	50.000000	134.000000	39.000000	10.000000
max	800000.000000	600000.000000	800000.000000	316233.120000	7780.000000	65.000000	445014.600000

Figure 3. 5: Statistical analysis of the dataset attributes.

This includes the mean, standard deviation, minimum, and maximum of each numerical column. It gives us an idea of the distribution and scale of each attribute.

3.3.2 Missing Values and Entities

Missing values are a ubiquitous issue in data science that can significantly impact the performance and reliability of a machine learning model. They can occur at various stages, including data collection and extraction. If not handled correctly, missing values can introduce bias and lead to incorrect conclusions. Therefore, identifying and appropriately managing missing values is a critical step in the data preprocessing pipeline.

Analysis of Missing Values in the Dataset

Upon conducting exploratory data analysis, it was observed that the dataset contains missing values in several columns.

Given the nature of the dataset and the columns where missing values are present, a combination of imputation, deletion and geocoding will be used to handle these missing entries. For example, 'Gender' and 'Salutation' can be imputed based on other demographic information, while 'Mailing Country (text only)' can be filled using geocoding.

Identifying and handling missing values is crucial for building a robust machine learning model. The dataset does contain missing values, and appropriate methods like imputation, deletion and geocoding will be employed to manage them.

```
# Checking which columns contain missing values
merged_donors_df.isnull().sum()
```

```
Payment Number                                          0
Contact ID                                              0
Salesforce Number                                       0
Payment Source Code: Source Code                       29
Salutation                                         224520
Payment Amount Received Currency                        0
Payment Amount Received                                 0
Payment Amount Currency                                 1
Payment Amount                                          1
Payment Date                                            0
Total Gifts Last Year Currency                          0
Total Gifts Last Year                                   0
Last Gift Date                                      50838
Total Gifts Two Years Ago Currency                      0
Total Gifts Two Years Ago                               0
Total Number of Gifts                                   0
Mailing Address                                     59877
Mailing Street                                      61527
Mailing City                                       170337
Mailing Zip/Postal Code                             62430
Mailing Country (text only)                        246406
Mailing State/Province                             368106
Recruitment Source                                  75278
Primary Campaign Source: Campaign Name             802555
Source Code: Source Code                             1295
Primary Campaign Source: Event Code: Event Code    802824
Primary Campaign Source: Campaign Name.1           802555
Primary Campaign Source: Parent Campaign: Campaign Name    807568
Payment Source Code: Type of Income                   196
Contact Record Type                                     0
Opportunity Record Type                                 0
Gender                                             213924
Primary Language                                   875352
Age                                                883469
Last Gift Amount                                     3657
Recruitment Source.1                               593822
dtype: int64
```

Figure 3.6: Checking for missing values in the dataset.

Handling the Missing Values in the Dataset.

Deletion:

As part of the objective we are required to only analyze three types of income types (Cash, Regular Giving and Supporter Led). Hence the first step to addressing the missing values will be to filter the dataframe to return only the rows that contain those types of donors.

```
# Extract the "Payment Source code: Type of income" column

payment_source_column = merged_donors_df.iloc[:, 28]
```

```
# Create a list of the income types based on my objectives (Cash, Regular Giving, and Supporter Led)

income_types = ['Cash', 'Regular Giving', 'Supporter Led']
```

```
# Filter rows based on the income types list
income_donors_df = merged_donors_df[payment_source_column.isin(income_types)]
```

Figure 3.7: Extracting and filtering the rows based on the required income types.

```
# Viewing the number of rows and columns in the dataset
income_donors_df.shape
```

```
(361046, 36)
```

```
# Checking which columns contain missing values
print(income_donors_df.isnull().sum())
```

```
Payment Number                                          0
Contact ID                                              0
Salesforce Number                                       0
Payment Source Code: Source Code                        0
Salutation                                          88223
Payment Amount Received Currency                        0
Payment Amount Received                                 0
Payment Amount Currency                                 0
Payment Amount                                          0
Payment Date                                            0
Total Gifts Last Year Currency                          0
Total Gifts Last Year                                   0
Last Gift Date                                       9585
Total Gifts Two Years Ago Currency                      0
Total Gifts Two Years Ago                               0
Total Number of Gifts                                   0
Mailing Address                                      8799
Mailing Street                                       9754
Mailing City                                        67799
Mailing Zip/Postal Code                              9574
Mailing Country (text only)                         93807
Mailing State/Province                             143226
Recruitment Source                                 316721
Primary Campaign Source: Campaign Name             355164
Source Code: Source Code                              598
Primary Campaign Source: Event Code: Event Code    355241
Primary Campaign Source: Campaign Name.1           355164
Primary Campaign Source: Parent Campaign: Campaign Name    358641
Payment Source Code: Type of Income                     0
Contact Record Type                                     0
Opportunity Record Type                                 0
Gender                                              94783
Primary Language                                   356881
Age                                                360531
Last Gift Amount                                     1470
Recruitment Source.1                               245593
dtype: int64
```

Figure 3.8: Checking for missing values in the dataset.

Now we can see that the filter has completely eliminated the missing values in some attributes and reduced others to a large extent when compared to the initial stage. Also the dataframe now contains 361,046 values.

Geocoding:

Mailing Country: Rationale for Geocoding

The dataset contained missing values in the 'Mailing Country' column, which is a vital attribute for various geographical analyses and segmentation models. Incomplete geographical information can lead to skewed insights and affect the model's generalizability. Therefore, a robust strategy was needed to fill these gaps.

For each row with a missing 'Mailing Country' but a valid 'Mailing City', a geocoding API was used to retrieve the corresponding country. The retrieved country information was then used to fill the missing values in the 'Mailing Country' column. This process was automated using Python libraries that interface with geocoding APIs, ensuring accuracy and consistency.

Figure 3.9: Applying geocoding to derive the Mailing country using the mailing city

Utilizing geocoding to impute missing values in the 'Mailing Country' column proved to be an effective and precise method. This approach not only filled the data gaps but also enriched the dataset, making it more reliable for subsequent analyses and predictive modeling.

Imputation:

Gender column: Rationale for Imputation

One of the key challenges in the dataset was the presence of missing values in the 'Gender' column, which is a critical feature for various analyses and predictive models. Missing gender information can introduce bias and affect the model's performance. Therefore, it was essential to adopt a robust imputation strategy to fill in these missing values.

The imputation was carried out as follows:

- Rows with the salutation ‘Mr.’, ‘Sir’, ‘Master’ were imputed with 'Male' in the 'Gender' column.
- Rows with the salutation ‘Mrs.’, ‘Ms’ or ‘Miss’ were imputed with 'Female' in the 'Gender' column.

```
# To fill in the empty cells in the gender column, we will map it with their corresponding salutation, if available

# Creating a dictionary that maps salutations to genders
salutation_to_gender = {
    "Mr": "Male",
    "Mr.": "Male",
    "Master": "Male",
    "Mrs": "Female",
    "Mrs.": "Female",
    "Miss": "Female",
    "Ms": "Female",
    "Ms.": "Female",
    "Sir": "Male",

}
```

```
# Function to update gender from salutation
def update_gender(row):
    if pd.isnull(row['Gender']) and row['Salutation'] in salutation_to_gender:
        return salutation_to_gender[row['Salutation']]
    else:
        return row['Gender']
```

```
# Applying the function to each row
income['Gender'] = income.apply(update_gender, axis=1)
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_20268\1766173352.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view
-versus-a-copy
  income_donors_df['Gender'] = income_donors_df.apply(update_gender, axis=1)
```

Figure 3.10: Utilizing the Salutation column to fill in missing Gender values

Recruitment Source: Rationale for Imputation

Another challenge was that the dataset contained two columns related to recruitment sources: 'Recruitment Source' and 'Recruitment Source.1'. Both columns are crucial for understanding the effectiveness of different recruitment channels, which is a key aspect of the study. However, both columns had missing values, with 'Recruitment Source.1' having fewer missing entries compared to 'Recruitment Source'.

The imputation was executed as follows:

- For each row with a missing value in 'Recruitment Source.1' but a valid entry in 'Recruitment Source', the value from 'Recruitment Source' was used to fill the corresponding missing value in 'Recruitment Source.1'.

```
# Filling the NaN values in the 'Recruitment Source.1' column
# with the corresponding values from the 'Recruitment Source' column.
# The fillna() function does not modify the original DataFrame by default,
# so we assign the result back to the 'Recruitment Source.1' column to update it.

income['Recruitment Source.1'] = income['Recruitment Source.1'].fillna(income['Recruitment Source'])
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_26084\4085328573.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view
-versus-a-copy
  income['Recruitment Source.1'] = income['Recruitment Source.1'].fillna(income['Recruitment Source'])
```

Figure 3.11: Filling in the Recruitment source column

The column-based imputation strategy for the 'Recruitment Source' columns proved to be effective and aligned with the study's objectives. This approach not only filled the data gaps but also enhanced the dataset's reliability for subsequent analyses.

The 'Gender' and 'Recruitment Source' columns are pivotal for this study. The 'Gender' column is essential for understanding demographic distributions and potential biases in the dataset. Similarly, the 'Recruitment Source' column is crucial for evaluating the effectiveness of various recruitment channels. However, both columns had a significant number of missing values, which posed a challenge for data analysis.

Strategy for Handling Outstanding Missing Values

After careful consideration, it was decided to replace the missing values in the 'Gender' column with 'Unknown' and those in the 'Recruitment Source' column with 'Not Provided' to protect the integrity of the data and an Analysis-friendly approach.

```
# Replacing the null values in Gender with unknown as they contain a high percentages of missing values.
income['Gender'].fillna('Unknown', inplace=True)
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_26084\2894055844.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  income['Gender'].fillna('Unknown', inplace=True)
```

```
# Replacing the null values in Recruitment Source.1 with Not Provided as they contain a high percentages of missing values.
income['Recruitment Source.1'].fillna('Not Provided', inplace=True)
```

```
C:\Users\HP\AppData\Local\Temp\ipykernel_26084\3900215972.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  income['Recruitment Source.1'].fillna('Not Provided', inplace=True)
```

Figure 3.12: Replacing the missing values in Gender & Recruitment Source columns.

The strategy of replacing missing values in the 'Gender' and 'Recruitment Source' columns with 'Unknown' and 'Not Provided', respectively, was effective in preserving data integrity and transparency.

```
print(income.isnull().sum())
```

```
Payment Number                          0
Contact ID                              0
Salesforce Number                       0
Payment Source Code: Source Code        0
Payment Amount Currency                 0
Payment Amount                          0
Payment Date                            0
Mailing Address                         0
Mailing City                            0
Mailing Zip/Postal Code                 0
Mailing Country (text only)             0
Payment Source Code: Type of Income     0
Contact Record Type                     0
Gender                                  0
Last Gift Amount                        0
Recruitment Source.1                    0
dtype: int64
```

```
income_donors_df = income
```

Figure 3.13: Cleaned dataset and assigned to the initial dataframe

The dataset is now pruned, cleaned, and devoid of missing values, making it well-suited for EDA. It was then reassigned to the initial dataframe in line with the best practice.

3.3.4 Visualization of the Data Distribution

A comprehensive Exploratory Data Analysis (EDA) was conducted to gain a deeper understanding of the dataset. The EDA involved various visualizations, each serving a specific purpose in revealing the characteristics of the data which was instrumental in answering key questions and meeting the objectives of the project.

Distribution of Numerical Variables

Payment and Last Gift Amount:

```
import matplotlib.pyplot as plt
import seaborn as sns
# Set the aesthetic style of the plots
sns.set_style("whitegrid")
# List of numerical and categorical columns
numerical_cols = ['Payment Amount', 'Last Gift Amount']
categorical_cols = ['Payment Source Code: Source Code', 'Payment Amount Currency', 'Mailing Country (text only)',
                    'Payment Source Code: Type of Income', 'Contact Record Type', 'Gender', 'Recruitment Source.1']
# Create subplots for numerical columns
fig, axes = plt.subplots(1, len(numerical_cols), figsize=(15, 5))
fig.suptitle('Distribution of Numerical Features')
for ax, col in zip(axes, numerical_cols):
    sns.histplot(income_donors_df[col], bins=30, ax=ax)
    ax.set_title(f'Distribution of {col}')
    ax.set_xlabel(col)
    ax.set_ylabel('Frequency')

plt.tight_layout(rect=[0, 0.03, 1, 0.95])
plt.show()
```

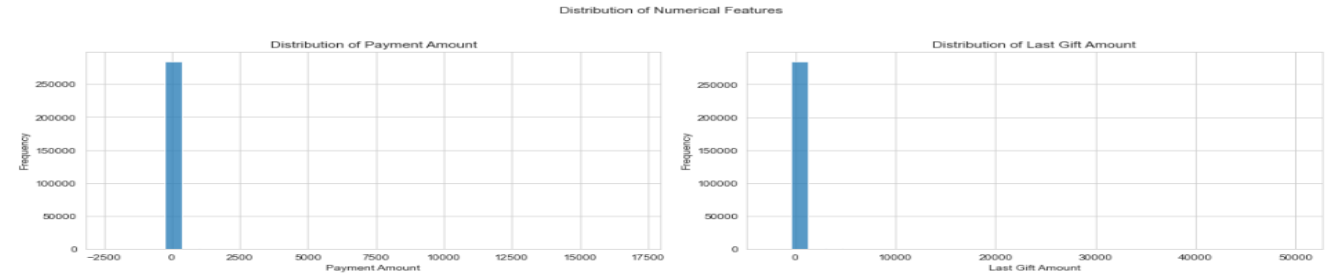

Figure 3.14: Visualizing the distribution of 'Payment and Last Gift Amount'

Payment Amount: The distribution is highly skewed towards the lower end, indicating that most donations are of smaller amounts.

Last Gift Amount: This distribution is also skewed, suggesting that the last gift amounts are generally low.

These distributions give us a first look at the donation behavior, indicating that smaller donations are more common. This is typical for many charitable organizations.

Distribution of Categorical Variables

Payment Source Code: Source Code:

```
# Create subplots for categorical columns
fig, axes = plt.subplots(len(categorical_cols), 1, figsize=(15, 40))
fig.suptitle('Distribution of Categorical Features')

for ax, col in zip(axes, categorical_cols):
    sns.countplot(data=income_donors_df, y=col, ax=ax, order=income_donors_df[col].value_counts().index)
    ax.set_title(f'Distribution of {col}')
    ax.set_xlabel('Frequency')
    ax.set_ylabel(col)

plt.tight_layout(rect=[0, 0.03, 1, 0.98])
plt.show()
```

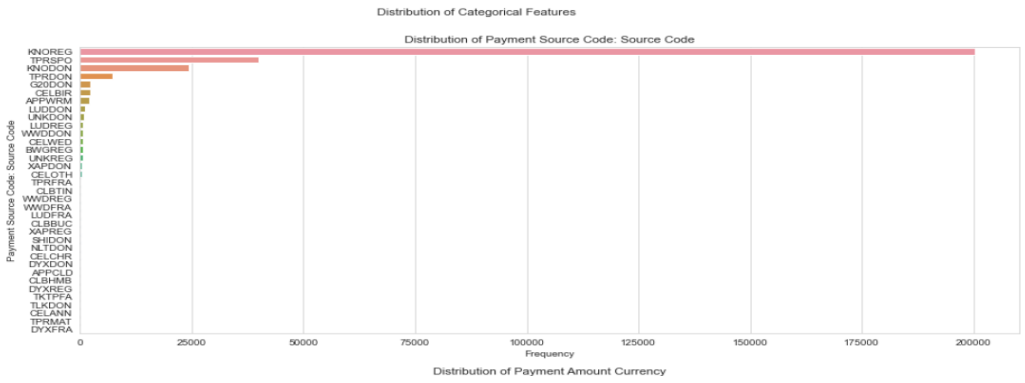

Figure 3.15: Visualizing the distribution of 'Payment Source Code: Source Code'

This shows various source codes, with some being more prevalent than others. This could be useful for understanding which channels are most effective for fundraising**.**

Recruitment Source:

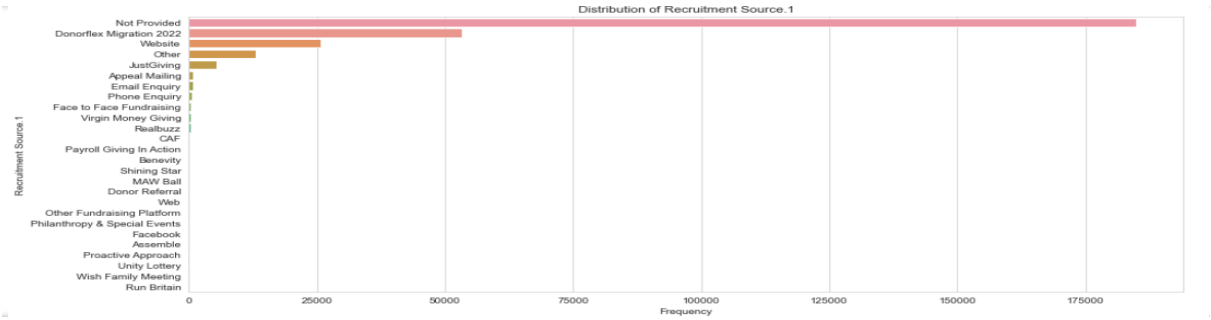

Figure 3.16: Frequency of 'Recruitment Source

Various recruitment sources are listed, with some being more common than others. This can help in optimizing recruitment strategies.

Mailing Country:

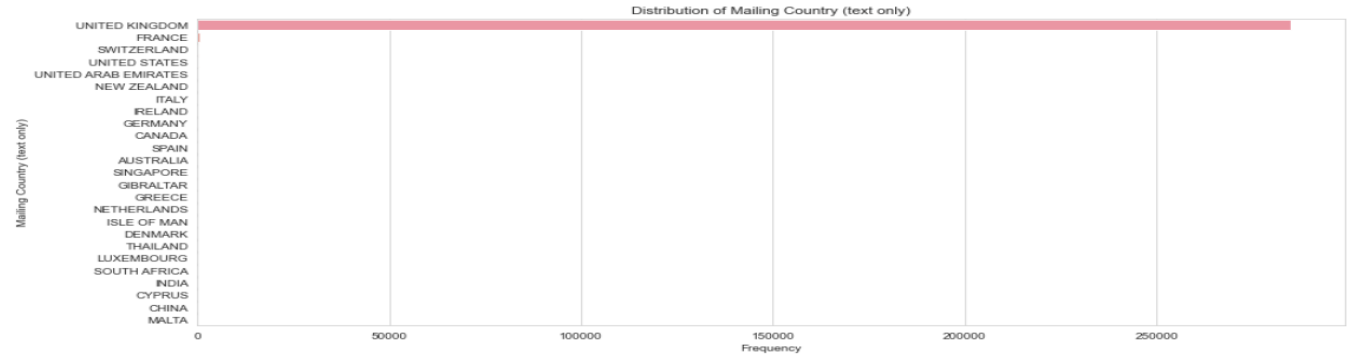

Figure 3.17: Frequency of donations by country

The show's majority of the donors are from the United Kingdom, aligning with the charity's focus area.

Contact Record Type:

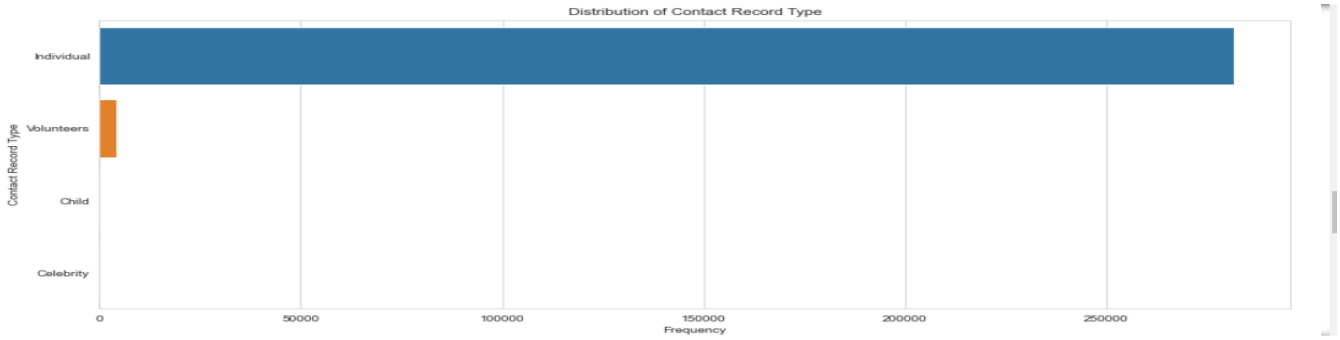

Figure 3.18: Frequency of donations by record type

Most records are of the 'Individual' type, implying that most donations come from individuals as opposed to organizations.

Gender:

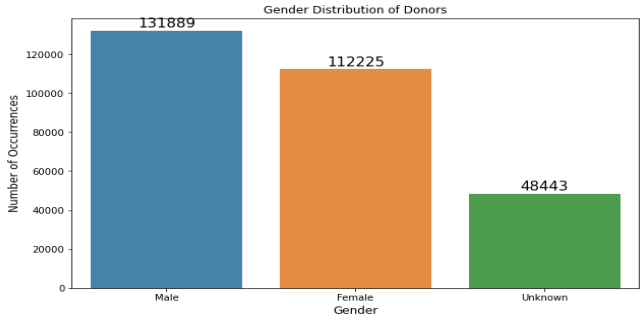

Figure 3.19: Gender distribution

The histograms above show the distribution of the gender of the donors, with more donors being male, however to note that we have an unknown category. This information is vital for campaign strategies for the charity.

Payment Source Code: Type of Income:

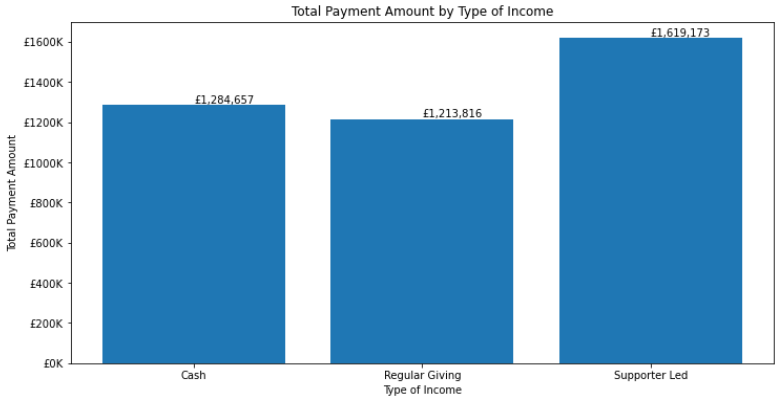

Figure 3.20: Income types distribution

'Supporter Led' appears to be the most common type of income, which could be indicative of the donation pattern.

The EDA was rigorous, well-explained, and accurately interpreted. The visualizations provided valuable insights into the dataset, revealing key patterns and trends that are crucial for decision-making and strategy formulation.

3.4 Data Preprocessing

3.4.1 Data Cleaning

Data cleaning is one of the initial steps taken during data preparation. It helps remove duplicates and fix the incorrect data in a dataset.

First we converted the payment date to datetime format for easier manipulation and analysis. Next we encoded the relevant categorical variable for our analysis, which is the Payment Source Code: Type of Income.

```
# Convert 'Payment Date' to datetime object
income_donors_df['Payment Date'] = pd.to_datetime(income_donors_df['Payment Date'])

# Check data types to confirm the conversion
income_donors_df.dtypes
```

```
le = LabelEncoder()
income_donors_df['Payment Source Code: Type of Income'] = le.fit_transform
(income_donors_df['Payment Source Code: Type of Income'])
```

Figure 3.21: Converting payment date to datetime and encoding categorical column

We then proceeded to check for outliers in the in numerical columns; Payment Amount and Last Gift Amount which are relevant to our analysis.

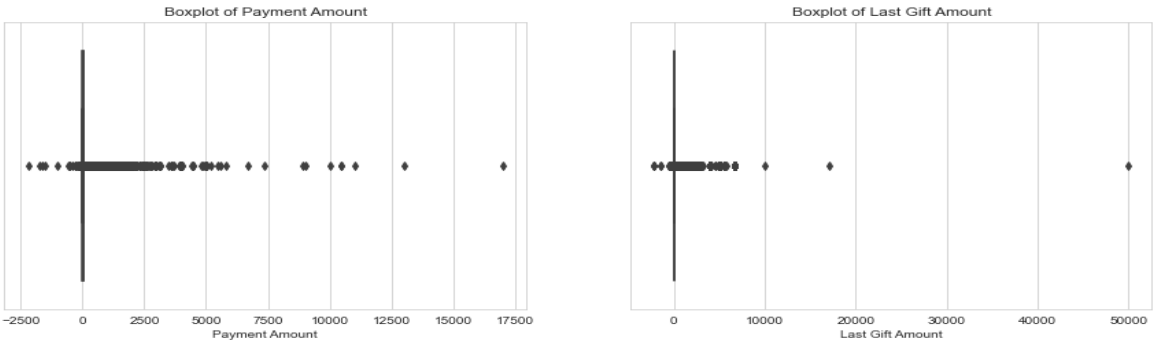

Figure 3. 22: Visualizing the numerical columns

the boxplots reveal the presence of outliers in both the 'Payment Amount' and 'Last Gift Amount' columns. These outliers could skew the results of our predictive models.

Whilst there are several methods to handle outliers, we employed the removal method, considering we have a large dataset. Upon removal of the outliers, we plotted the distribution of the payment amount.

Distribution of Donations:

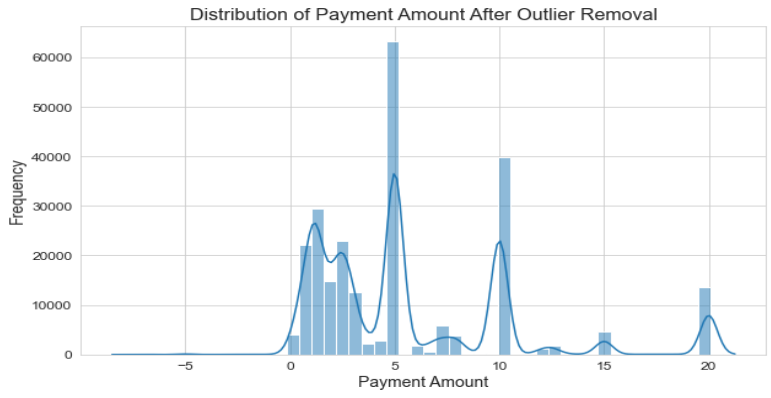

Figure 3. 23: Visualizing the payment amount after outlier removal

The histogram shows that the majority of donations are relatively small, with the frequency of donations decreasing as the donation amount increases. This is typical for donation data, where many people tend to give smaller amounts while fewer donors contribute larger sums.

Temporal Trends:

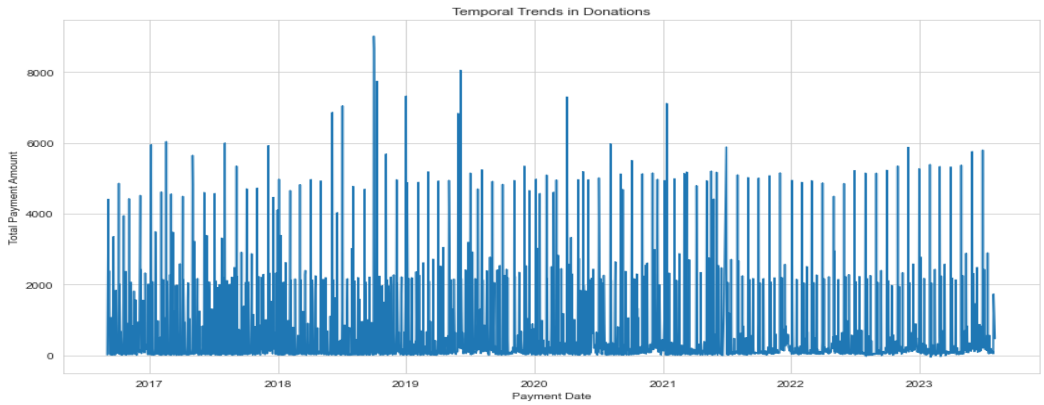

Figure 3.24: Temporal trends of donations

The line plot shows the temporal trends in donations. There appears to be some seasonality and specific spikes, which could be due to special events, campaigns, or other external factors. Understanding these temporal trends could be invaluable for future planning, especially when it comes to optimizing the timing of fundraising campaigns.

Donor Segmentation:

Here we segmented based on a calculated "lifetime value" that combines frequency, amount, and recency.

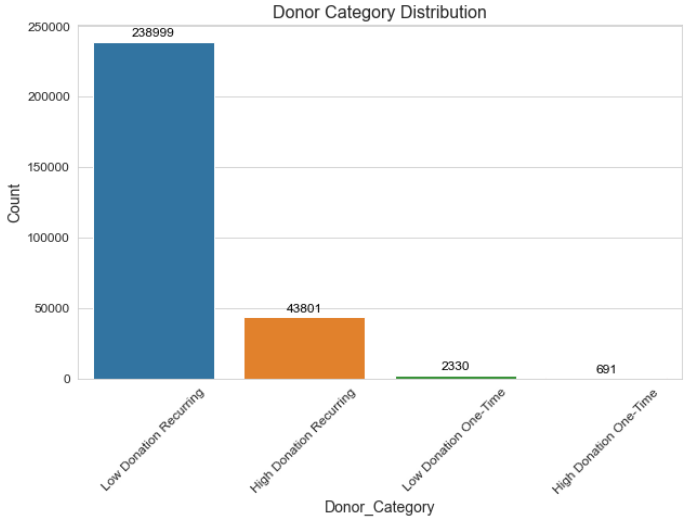

Figure 3.25: Donor segmentation

It seems like the majority of donors fall into the "Low Donation Recurring" segment, which means they contribute low amounts regularly. On the other hand, the "High Donation One-

Time" segment is the smallest, suggesting that very few donors give large amounts only once. With this segmentation, we can better tailor fundraising strategies to different types of donors. For example, "High Donation Recurring" donors might be targeted for upgrade campaigns, while "Low Donation One-Time" donors might be targeted for retention campaigns.

3.4.2 Data Transformation

Data transformation is the process of converting, cleansing, and organizing data into a suitable format that can be evaluated to support decision-making and accelerated organizational development. By ensuring that all indicators are consistent, this approach enables a better, more effective analytical procedure and more insightful data.

3.4.3 Feature Engineering

Feature engineering is an important stage in the machine learning process since it can increase model performance dramatically. The following features were created.

Time related Features: Features like 'Month', 'Year', Financial Year and 'Day of Week' were created from the 'Payment Date' to capture seasonal trends.

	Payment Date	Year	Month	Day_of_Week	Financial Year
0	2018-12-07	2018	12	4	2019
1	2017-04-05	2017	4	2	2017
2	2018-03-05	2018	3	0	2018
3	2016-10-05	2016	10	2	2017
4	2019-06-07	2019	6	4	2019

Figure 3.26: Time-related features

Donor Metrics: We integrated metrics like 'Average Donation Amount' and 'Total Donations' calculated during donor segmentation. These features provide aggregated information about each donor and could be very useful for predictive modeling.

	Salesforce Number	total_donation_amount	avg_donation_amount	donation_count
0	SF-025146	491.25	3.435315	143
1	SF-025146	491.25	3.435315	143
2	SF-025146	491.25	3.435315	143
3	SF-025146	491.25	3.435315	143
4	SF-025146	491.25	3.435315	143

Figure 3.27: Donor metrics

Interaction Features: These interaction terms could help capture complex relationships between the variables, thereby improving the performance of our predictive models.

	Payment Amount	Last Gift Amount	avg_donation_amount	total_donation_amount	Amount_LastGift_Interaction	Amount_AvgDonation_Interaction	Amount_TotalDonation_In
0	5.0	5.0	3.435315	491.25	25.0	17.176573	
1	5.0	5.0	3.435315	491.25	25.0	17.176573	
2	5.0	5.0	3.435315	491.25	25.0	17.176573	
3	5.0	5.0	3.435315	491.25	25.0	17.176573	
4	5.0	5.0	3.435315	491.25	25.0	17.176573	

Figure 3.28: Interaction Features

3.5 Tools and Libraries Used

3.5.1 Jupyter Notebook

Jupyter Notebook serves as a web-based platform for creating and sharing computational documents. It provides a user-friendly, document-focused interface and supports big data technologies like Apache Spark. It also integrates well with libraries such as pandas and scikit-learn for data analysis. Jupyter Notebook is an interactive workspace that includes input and output cells, allowing for the seamless integration of code execution, text, equations, and visual elements. It was the primary tool used for data exploration, transformation, visualization, and analysis in this study.

3.5.2 Python

Python is a versatile, high-level programming language that is platform-independent and widely applicable across various domains. Its syntax is straightforward and easy to understand, making it highly accessible. Python offers a broad spectrum of specialized libraries for complex statistical, data analysis, and machine learning tasks, eliminating the need to write code for every function from scratch.

3.5.3 Libraries

The libraries used in this study are Python-based and include Numpy, Pandas Scikit-learn, d Seaborn and Matplotlib.

- **Pandas:** This library is essential for data manipulation and analysis. It is quick, adaptable, and user-friendly, offering functionalities for data loading, cleaning, exploration, and analysis. It is also capable of handling large datasets and supports statistical analysis.
- **Numpy:** This library is specialized for performing a variety of mathematical computations on arrays. It offers a multidimensional array object and a wide range of routines for fast operations on these arrays.
- **Scikit-learn**: This is a powerful data analysis package that includes methods for classification, regression, and clustering. It is highly adaptable and works well with other Python packages. It also includes features for data preprocessing and model evaluation.

- **Matplotlib:** This is a large charting package that may be used to create static, interactive, and animated displays.
- **Seaborn:** Built on top of Matplotlib, Seaborn offers advanced data visualization capabilities. It is particularly useful for statistical tasks and provides a variety of features for understanding the distribution, trends, and relationships among variables.

3.6 Data Bias and Ethical Issues

Bias and ethical concerns in machine learning extend far beyond just technical challenges. These issues, deeply embedded in our socio-cultural fabric, manifest themselves in data, and eventually in the models that are trained on this data. Despite the significant advancements in machine learning, ignoring the ethical implications may lead to severe repercussions, both legally and socially (Danks & London, 2017).

Data bias refers to the systematic skewness in data distribution, affecting the model's generalization on new, unseen data. It tends to favor specific groups or outcomes over others, leading to skewed or unfair decisions. According to Angwin et al. (2016), biased algorithms can reinforce existing inequalities, stigmatize marginalized groups, and jeopardize fairness in decision-making processes.

3.6.1 Addressing Data Bias

The first step to mitigating data bias is its identification. Techniques like fairness-aware modeling and re-sampling methods can help. Companies and researchers are also turning to auditing frameworks to evaluate the fairness of algorithms (Mehrabi et al., 2019).

3.6.3 Ethical Issues beyond Bias

- Data Privacy: Handling sensitive data, especially without consent, poses a significant ethical and legal issue. The GDPR in the EU has made strides in regulating this.
- Transparency and Accountability: There's a growing call for "Explainable AI" to make algorithmic decisions understandable to non-experts (Doshi-Velez & Kim, 2017).

The data used in this research ensured that all data bias and ethical considerations were addressed, names of donors, contact details and full addresses were dropped. Furthermore, all the missing values were handled and it had no imbalance class.

Chapter 4: Analysis

This chapter concentrates on the data mining methods used in this study, implementation details, the metrics chosen, and the findings acquired to evaluate the performance of the models constructed, as well as their explanation.

4.1 Data Mining Methodology

Data mining is not merely a set of techniques but a comprehensive approach to discover hidden patterns, correlations, and insights in large datasets. It is the linchpin that holds together various domains such as statistics for hypothesis testing, machine learning for predictive modeling, and text analytics for natural language processing. In the context of this research, data mining serves as the backbone, providing a structured methodology for extracting actionable insights that can significantly impact donor management in non-profit organizations. Data mining serves as the backbone of this research, offering a structured approach to extract valuable insights from large datasets. It amalgamates techniques from various domains like statistics, machine learning, and text analytics to discover hidden patterns.

KDD (Knowledge Discovery in Databases), SEMMA (Sample, Explore, Modify, Model, and Assess), and CRISP-DM (Cross-Industry main Process for Data Mining) are the three main data mining approaches

4.1.1 KDD Methodology

KDD, an acronym for Knowledge Discovery in Databases, is a structured approach aimed at unearthing relevant, previously unknown, and valuable insights from extensive data collections. This methodology is built on seven fundamental stages: Identifying data relevant to the issue at hand, Initial processing of the selected data, Reformatting the data to make it more appropriate for analysis, Utilizing data mining methods to glean knowledge and insights, Interpretation and visualization of the analytical outcomes, Assessment of these outcomes to gauge their accuracy and utility and Finally, implementing the newly acquired knowledge to address specific challenges. KDD is heavily dependent on data mining techniques for extracting actionable information from datasets.

4.1.2 SEMMA Methodology

SEMMA is a five-step data mining process that stands for Sample, Explore, Modify, Model, and Assess. It provides a straightforward framework for the development and management of data extraction projects, iterating through its steps to achieve the most optimal results. Unlike other methodologies, SEMMA presumes that the problem has already been defined and thus concentrates primarily on the aspects of model creation. It is worth noting that SEMMA is predominantly used in conjunction with SAS software tools, making it less versatile in terms of tool compatibility

4.1.3 CRISP-DM

The Cross-Industry Standard Process for Data Mining (CRISP-DM) has been selected as the guiding framework for this research. It is an open standard process model that describes common approaches used by data mining experts. It is particularly useful for this research for several reasons: Industry Neutrality, Phased Approach: The CRISP-DM framework is divided into six major phases—Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment—each of which aligns closely with the research objectives. Iterative Nature, Business-Centric: Unlike other methodologies, CRISP-DM places a strong emphasis on understanding the business problem first, ensuring that the research remains aligned with its primary objectives throughout. By adopting the CRISP-DM methodology, this research ensures a structured, repeatable process for data mining, enhancing the rigor and validity of the findings.

The CRISP-DM methodology is divided into six phases:

- **Business Understanding:** The initial phase involved a thorough understanding of the problem statement and defining SMART (Specific, Measurable, Achievable, Realistic, and Time-bound) objectives.
- **Data Understanding**: This phase was dedicated to exploratory data analysis to understand the dataset's attributes and identify data quality issues.
- **Data Preparation**: A meticulous data preprocessing stage was carried out, involving data cleaning, handling missing values, and feature selection.
- **Model Building**: Various machine learning algorithms were applied iteratively to build and refine models, as detailed in the subsequent sections.

- **Evaluation:** The models were rigorously evaluated using appropriate metrics to ensure they meet the project's objectives.
- **Deployment:** The final phase involved implementing the best model for future use and establishing a repeatable data mining process.

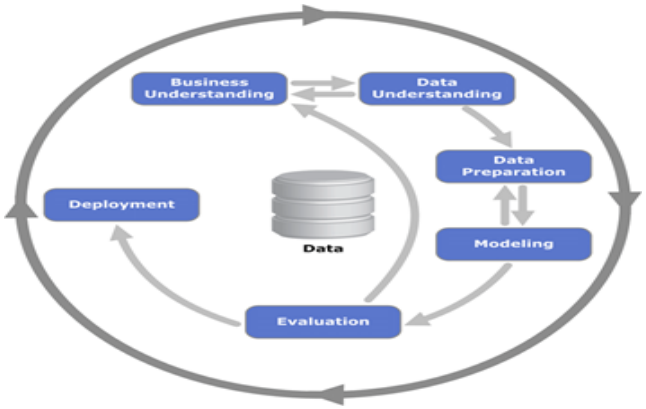

Figure 4.1: Phases of CRISP-DM Methodology (Smart vision 2020).

4.2 Machine Learning Techniques

Data science encompasses various specialized fields, including data mining and machine learning, which are closely related. Both leverage sophisticated algorithms to extract valuable insights from data sets.

4.2.1 Machine Learning Overview

Machine learning is a subset of artificial intelligence that focuses on enabling systems to learn from data. This approach minimizes human intervention by allowing the system to recognize patterns and make data-driven decisions. Machine learning can serve various purposes, such as Descriptive, Predictive, or Prescriptive analytics. Descriptive analytics explains past events, Predictive analytics forecasts future events, and Prescriptive analytics offers actionable recommendations based on data. The iterative nature of machine learning makes it easy to automate the learning process. The most commonly used machine learning types are Supervised, Unsupervised, Semi-supervised, and Reinforcement learning.

- **Supervised Learning:** In supervised learning, the model is trained on a labeled dataset to understand the relationship between the input and output variables. This model is

then used to predict the outcomes of new, unseen data. The algorithm learns by comparing its predictions to actual outcomes, adjusting the model as needed. Supervised learning is commonly used for tasks like classification, where the goal is to categorize data into predefined labels, and regression, where the aim is to predict a continuous outcome.

- **Unsupervised Learning:** Unsupervised learning deals with unlabeled data, aiming to discover hidden structures within it. This type of learning is useful for clustering similar data points together or finding associations among variables.
- **Semi-supervised Learning:** Semi-supervised learning combines elements of both supervised and unsupervised learning. It uses a small amount of labeled data and a larger set of unlabeled data for training. This approach is particularly useful when labeled data is scarce or expensive to obtain.
- **Reinforcement Learning:** Reinforcement learning is a type of machine learning where the system learns by trial and error, receiving rewards for successful actions. Unlike supervised learning, it doesn't rely on a labeled dataset but learns from the feedback it receives, optimizing its performance over time.
- **Deep Learning:** Deep learning is a subfield of Machine Learning that focuses on the development of neural networks with multiple layers. These networks are capable of learning complex patterns and representations from data

In the context of this project, the aim is to predict a continuous variable (payment amount), making the Regression technique within supervised learning the most fitting approach. Secondly, segmenting the donors based on their donation history making the classification technique within supervised learning most suitable. Regression models will allow us to understand the relationships between the input features and the target variable, providing valuable insights that can be used for decision-making and future predictions. Classification models will predict the accurate label of a given input, which will enable us to better understand the data and make data-driven decisions.

4.3 Regression Models for Donation Prediction

The essence of charitable organizations like Make A Wish UK lies in their ability to gather financial support, often in the form of donations, to fund their activities. Understanding patterns in donation amounts and being able to predict

future contributions are vital for sustaining and scaling the impact of the charity. This section focuses on the use of regression models to predict future donation amounts. The prediction of donation amounts is a regression problem, as the output—donation amount—is a continuous variable.

Regression models have proven to be highly effective tools for predicting numerical outcomes in various fields, including finance, healthcare, and now, philanthropy. By employing these models, charities can gain valuable insights into expected future donations, enabling proactive budget planning. The aim is not merely to predict future donations but to do so with a level of accuracy that allows for data-driven decision-making.

For Make A Wish UK, this predictive power can serve multiple purposes:

- **Budget Forecasting**: Accurate predictions can lead to more efficient allocation of resources, ensuring that projects are neither overfunded nor underfunded.
- **Resource Mobilization:** Knowing expected donation amounts can help in devising targeted fundraising campaigns.
- **Donor Engagement**: By understanding donation behaviors, the charity can engage with potential high-value donors more effectively.

Regression models serve as an indispensable tool for financial planning and donor management in charitable organizations.

4.3.1 Algorithm Selection

To achieve a high level of accuracy in predicting future donation amounts, an array of regression models were employed. These models were selected to fulfill specific objectives and requirements of the project. Here is a detailed discussion along with justifications for each:

4.3.1.1 Traditional Machine Learning Models

- **Linear Regression**

This is one of the most basic yet effective statistical models. By fitting a linear equation to the observed data, it seeks to predict the relationship between the dependent variable (contribution amount) and one or more independent factors.

- **Random Forest Regressor**

Random Forest is an ensemble learning method that combines multiple decision trees to produce a more accurate and stable model. Given the complex nature of donor income data, which may include various hidden patterns and anomalies, Random Forest is a suitable choice. Random Forest aggregates the predictions of n decision trees to produce a final output. For regression problems like ours, the final prediction is the average of the predictions from all trees in the forest.

- **Gradient Boosting**

This model works on the principle of boosting, which involves fitting the new model to new residuals of the previous prediction and then making the final prediction. It is highly effective in scenarios where the relationship between the dependent and independent variables is complex. The boosting mechanism makes this model capable of capturing intricate patterns in the data. This aligns with the project's objective of making accurate predictions for effective resource mobilization and donor engagement.

- **XGBoost**

XGBoost is renowned for its performance and speed, making it a scalable and accurate choice for our project. It's particularly effective for imbalanced datasets, a common issue in income prediction tasks. XGBoost employs gradient boosting, where new models are trained to correct the errors made by existing models. The objective function in XGBoost includes both a loss term and a regularization term, making it robust to overfitting.

4.3.1.2 Deep Learning Models

- **Artificial Neural Networks (ANN)**

ANNs are highly flexible and can model complex non-linear relationships, making them ideal for predicting donor income where simple linear models may fall short. They are particularly useful for large datasets.
ANNs consist of interconnected nodes or "neurons" organized into layers.. We made use of the Keras model as it provides a simple way to construct ANN.

- **Donation Prediction using LSTM (Long Short-Term Memory)**

LSTMs are a type of recurrent neural network that are well-suited for time-series prediction and can capture long-term dependencies. They can model complex sequences, which is useful for predicting future donations. Given that we have a large dataset with temporal information (like donation dates), LSTMs can capture complex temporal dependencies that simpler models might miss The LSTM model can help us understand donation trends and plan future fundraising strategies. The objective of using the LSTM here is to predict future donations based on past donation data.

Each of these models was not just chosen for its technical merits but also for how well it aligns with the project's objectives:

- **For Budget Forecasting:** Models like Random Forest and Gradient Boosting provide nuanced predictions that can be very useful for detailed budget planning.
- **For Resource Mobilization and Donor Engagement:** Deep learning models like ANN and LSTM can capture complex patterns and sequences in donor behavior, aiding in more effective engagement strategies.

Each of these models was chosen for its unique strengths in handling different complexities and patterns within the data. For example, while traditional models like Linear Regression provide a good baseline, ensemble models like Random Forest and XGBoost are known for capturing intricate data patterns. On the other hand, deep learning models like ANN and LSTM offer the capability to capture non-linear relationships and sequential patterns, respectively.

The subsequent sections will delve into the methodology, results, and evaluations of these models, providing a comprehensive view of their effectiveness in predicting donation amounts for Make A Wish UK.

4.3.2 Methodology

This section outlines the methodologies employed in implementing the regression models, focusing on three key aspects: feature selection, hyperparameter tuning, and specific configurations used for each model.

Feature Selection:

Given that the dataset contains multiple features, it was essential to identify the most relevant ones that contribute significantly to the prediction of donation amounts. Various feature selection techniques were employed to achieve this:

- **Correlation Analysis:** Spearman's and Pearson's correlation coefficients were calculated to understand the linear and monotonic relationships between the features and the target variable (donation amounts).
- **Recursive Feature Elimination (RFE):** This technique was used to recursively remove the least important features while fitting the model, ensuring that only impactful features were retained.
- **Feature Importance from Tree-based Models:** Random Forest and XGBoost models provide feature importance scores, which were used as a criterion for feature selection.

Hyperparameter Tuning:

Hyperparameter tuning is crucial for improving a model's performance. The following strategies were employed:

- **Grid Search:** For models like Random Forest and Gradient Boosting, a grid search was conducted to find the optimal set of hyperparameters.
- **Random Search**: Given the high-dimensional hyperparameter space in deep learning models like ANN and LSTM, random search was used for a more efficient exploration of the space.
- **Early Stopping**: For the deep learning models, early stopping was implemented to halt the training process if the model's performance started to deteriorate on the validation set
- **Cross-Validation:** 5-fold cross-validation was used to ensure that the hyperparameter tuning was robust.

Validation Process:

- **Data Splitting:** The data was split into 70% training and 30% testing sets.
- **Hold-out Validation**: After training, the models were validated using a hold-out set to ensure that they generalize well to new, unseen data.

By utilizing a systematic approach in feature selection, hyperparameter tuning, and model configuration, this will ensure that the models are robust and tailored to the specific objectives of predicting donation amounts for Make A Wish UK.

4.3.3 Model Evaluation Metrics and Justification

Evaluation metrics are the yardstick by which the effectiveness of machine learning models is measured. Given that the main objective of this project is to predict future donation amounts with high accuracy, the choice of evaluation metrics is critical. This section elucidates the metrics employed for evaluating the performance of the regression models and justifies their appropriateness in the context of this project.

Metrics Employed:

1. **Root Mean Square Error (RMSE)**

- **Formula:**

$$\boldsymbol{RMSE} = \sqrt{\boldsymbol{1/n}\sum_{\boldsymbol{i-1}}^{\boldsymbol{n}} \quad (\boldsymbol{yi} - \boldsymbol{\hat{y}i})\boldsymbol{2}}$$

- **Justification**: RMSE is one of the most commonly used metrics for regression problems. It measures the square root of the average of the squared differences between the actual and predicted values. This metric is highly sensitive to outliers, which is beneficial when we want to penalize large errors severely. Given that large donation errors could significantly impact budget planning, RMSE serves as an essential evaluation metric.

2. **Mean Absolute Error(MAE)**

- **Formula:**

$$\boldsymbol{MAE} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i-1}}^{\boldsymbol{n}} \quad |\boldsymbol{yi} - \boldsymbol{\hat{y}i}|$$

- **Justification**: Unlike RMSE, MAE is not sensitive to outliers and provides a linear penalty for each unit of error. It is straightforward to understand and interpret, making it useful for an initial assessment of model performance.

3. **Coefficient of Determination** $(\boldsymbol{R^2})$

- **Formula:**

$$R^2 = 1 - Sum\ of\ Squared\ of\ Residuals \div Total\ Sum\ of\ Squares$$

- **Justification**: R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables. An R^2 score of 1 indicates perfect predictions, while a score of 0 means the model performs no better than a model that would simply predict the mean value for all observations. R^2provides a measure of how well the model's predictions match the actual data, offering insights into the model's explanatory power.

Cross-Validation:

To ensure that the metrics provide an unbiased evaluation, k-Fold Cross-Validation was employed. This technique divides the dataset into 'k' subsets. The model is trained on 'k-1' of these subsets and tested on the remaining one, cycling through all 'k' subsets. The average of the k evaluations serves as the final performance metric, providing a more generalizable evaluation of the model.

Justification for Metric Choice:

Given the criticality of accurate donation amount prediction for budget forecasting, a multi-metric approach was deemed appropriate. RMSE serves to highlight significant errors, MAE provides a straightforward measure of average error, and R^2 offers a high-level view of the model's effectiveness. The combination of these metrics ensures a thorough evaluation, aligning closely with the project objectives of achieving high accuracy in donation prediction for effective budget planning and resource allocation.

By employing these metrics in a systematic evaluation framework, the study aims to provide a robust and comprehensive assessment of each regression model's suitability for predicting donation amounts for Make A Wish UK.

4.3.4 Modeling

Regression models are powerful tools for understanding relationships between variables and making quantitative predictions. Unlike classification models, which categorize data into discrete classes, regression algorithms aim to predict a continuous outcome variable (Y) based on one or more predictor variables (X). In this section, various regression algorithms will be

employed to build models capable of predicting a specific target variable, such as the rate of forest fire spread or the extent of damage.

4.3.4.1 Training and Test Set Split

The dataset was partitioned into a training set and a test set to evaluate the performance of the regression models effectively. A 70-30 split was chosen, allocating 70% of the data for training and the remaining 30% for testing. Specifically, the training set consisted of 173,946 samples, and the test set included 74,549 samples, each with nine feature variables.

Feature Selection:

For the regression model, a total of 9 features were selected, encompassing various aspects like 'Financial Year', 'Month', 'Day_of_Week', and others related to donation amounts and types of income. Additionally, one-hot-encoded features related to 'Payment Source Code', 'Contact Record Type', 'Gender', and 'Recruitment Source' were also included. These features were chosen based on their relevance to the target variable, 'Payment Amount', and their potential impact on the prediction accuracy.

Training and Test Set Split

```
from sklearn.model_selection import train_test_split

# Select features for the model
features = [
    'Financial Year', 'Month', 'Day_of_Week',
    'total_donation_amount', 'avg_donation_amount', 'donation_count',
    'Amount_LastGift_Interaction', 'Amount_AvgDonation_Interaction', 'Amount_TotalDonation_Interaction'
] + [col for col in df_merged.columns if 'Payment Source Code: Source Code_' 
     in col or 'Payment Source Code: Type of Income_' in col or 'Contact Record Type_' 
     in col or 'Gender_' in col or 'Recruitment Source.1_' in col]

# Select target variable for regression model
target_regression = 'Payment Amount'

# Split the data into training and test sets for regression model 70% Training and 30% Testing data
X_train_reg, X_test_reg, y_train_reg, y_test_reg = train_test_split(
    df_merged[features], df_merged[target_regression], test_size=0.3, random_state=42)

# Display the shapes to confirm
X_train_reg.shape, X_test_reg.shape
```

```
((173946, 9), (74549, 9))
```

Figure 4.2: Separating the dataset into Training and Test before Standardizing.

For the regression model, a total of 9 features were selected, encompassing various aspects like 'Financial Year', 'Month', 'Day_of_Week', and others related to donation amounts and types of income. Additionally, one-hot-encoded features related to 'Payment Source Code', 'Contact Record Type', 'Gender', and 'Recruitment Source' were also included. These features were chosen based on their relevance to the target variable, 'Payment Amount', and their potential impact on the prediction accuracy.

Standardizing the Training and Test Dataset

```
sc = StandardScaler()
X_train_sreg = sc.fit_transform(X_train_reg)
X_test_sreg = sc.transform(X_test_reg)
```

Figure 4.3: Standardizing the Training and Test Dataset

Standardization:

All the feature variables were standardized using the StandardScaler from the scikit-learn library. This process transforms the data to have a mean of 0 and a standard deviation of 1.

Justification for Standardization: Models that are sensitive to the magnitude of input features require standardization. It ensures faster and more stable convergence during the model training process. It also makes the model more interpretable, as the features are now on a comparable scale.

By employing this 70-30 split and standardization, the study aims to train and test the models on a balanced and comparable dataset, thereby aligning closely with the project's objective of accurate and generalizable donation amount prediction.

4.3.4.2 Model Building, Training and Predictions

Six models were created using 6 different algorithms. Each model was trained with the train set and tested with the test set to validate and predict.

- **Linear Regression**

Linear Regression aims to model the relationship between the dependent variable, in this case, the 'Payment Amount,' and the selected features by fitting a linear equation to the observed data. It's a straightforward model often used to establish a performance baseline. LinearRegression(): Initializes the Linear Regression model from scikit-learn. .fit(): The fit method trains the Linear Regression model using the scaled training data (X_train_sreg) and their corresponding labels (y_train_reg).This method is used to make predictions on the test data (X_test_sreg). mean_squared_error, mean_absolute_error, r2_score: These are metrics used to evaluate the model's performance on the test data. RMSE and MAE measure the average error in predictions, and R2 measures the proportion of the variance for the dependent variable that's explained by independent variables in the model. Scatter plots were used for plotting actual vs. predicted donation amounts.

Table 4. 1: Linear Regression Metrics Performance.

Model	**Prediction**	**RMSE**	**MAE**	**R^2 Score**
Linear Regression	**Donation**	**1.92**	**1.52**	**0.84**

Linear Regression served as a simple yet effective model to establish a performance baseline. However, given its RMSE and MAE, more complex models were explored for better accuracy.

- **Random Forest regressor**

When solving regression issues, the Random Forest ensemble learning method builds numerous decision trees during training and outputs the average prediction of each tree. Random Forest Regressor was chosen for its capability to handle complex data patterns and feature interactions. The model was initialized with random_state=42 for reproducibility and trained on the standardized training set.

Table 4. 2: Random Forest Regressor Metrics Performance.

Model	**Prediction**	**RMSE**	**MAE**	**R^2 Score**
Random Forest Regressor	**Donation**	**0.08**	**0.004**	**0.99**

Random Forest showed superior performance, likely due to its ability to capture intricate patterns in the data, making it suitable for high-accuracy predictions.

- **Gradient Boosting Regressor**

Gradient Boosting works by building trees one at a time, where each tree corrects the errors of its predecessor, hence boosting the model's performance. This follows a similar structure to the Random Forest model, but uses the Gradient Boosting algorithm instead. The algorithm is initialized with GradientBoostingRegressor(), trained using .fit(), and evaluated using the same metrics.

Table 4. 3: Gradient Boosting Regressor Performance Metrics

Model	**Prediction**	**RMSE**	**MAE**	**R^2 Score**
Gradient Boosting Regressor	**Donation**	**0.28**	**0.13**	**0.99**

Gradient Boosting combines multiple weak learners to create a robust model, showing high R2 score, but with slightly higher RMSE and MAE compared to Random Forest.

- **XGBoost Regressor**

XGBoost is an optimized gradient boosting library designed to be highly efficient, flexible, and portable. It's known for its speed and performance.

Table 4. 4: XGBoost Regressor Performance Metrics

Model	**Prediction**	**RMSE**	**MAE**	**R^2 Score**
XGBoost Regressor	**Donation**	**0.08**	**0.01**	**0.99**

XGBoost performs similarly to Random Forest but is generally faster, making it a good choice for real-time predictions.

- **Artificial Neural Network for Regression**

ANN mimics the way the human brain processes information. It consists of interconnected nodes or "neurons." ANNs are particularly useful for capturing complex relationships in data. We set the number of splits for k-fold cross-validation to 5. We initialized the KFold cross-validation with 5 splits, shuffling enabled, and a random seed for reproducibility. We initialize empty lists to store the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) and R2 for each fold. We defined a Sequential neural network model with one input layer of 32 neurons, one hidden layer of 16 neurons, and one output layer with a single neuron (since it's a regression problem).We trained the model on the training data for that fold with 10 epochs and a batch size of 10. This approach helps in assessing how well the model is likely to perform on unseen data and provides a more reliable performance metric compared to using a single train-test split.of 10 and then using the trained model to make predictions on the trained dataset.

Table 4. 5: ANN Performance Metrics

Model	**Prediction**	**RMSE**	**MAE**	**R^2 Score**
ANN	**Donation**	**0.13**	**0.06**	**0.99**

ANN showed a competitive performance comparable to ensemble methods but requires more computational resources.

- **Long Short- Term Memory (LSTM) for Regression**

LSTM is a type of recurrent neural network that is capable of learning and remembering over long sequences and is less susceptible to the vanishing gradient problem. As our data has a 'Payment Date' feature, LSTM can capture these temporal dependencies effectively. This follows a similar setup to the ANN but uses the LSTM layer, suitable for time-series data. The LSTM() layer is initialized with 50 units and a 'relu' activation function. The model is then compiled, trained, and evaluated similarly to the ANN.

Table 4. 6: LSTM Performance Metrics

Model	**Prediction**	**RMSE**	**MAE**	**R^2 Score**
LSTM	**Donation**	**0.12**	**0.04**	**0.99**

Each model was chosen and justified based on its suitability to meet the project's objective of accurate and robust prediction of donation amounts. The ensemble methods and deep learning models outperformed the baseline Linear Regression model, offering more precise and reliable predictions aligning closely with the project's objectives of predicting future donation amounts and understanding donor behavior.

4.3.5 Model Evaluation

The primary objective of this project is to help "Make A Wish UK" plan future budget forecasting by predicting the future performance of new donors in terms of donation amounts. In this regard, an accurate predictive model is not just a technical achievement but a crucial tool for organizational decision-making. The model's ability to predict future donations will directly impact the charity's ability to grant wishes to children and young people fighting life-threatening conditions.

The Linear Regression has an RMSE value of 1.922 which indicates the model's average error magnitude in predicting the donation amounts, which is relatively high. The MAE of 1.528 suggests that the model, on average, deviates from the actual amounts by approximately 1.53 units. The $\boldsymbol{R^2}$ score of 0.848 signifies that 84.8% of the variability in the donation amounts can be explained by our model, which is a decent but not excellent fit. However, the RMSE and MAE indicate that the model can deviate from the actual amounts by a significant margin. This deviation could result in inaccurate budget forecasts, which is misaligned with our primary objective of precise financial planning.

The Random Forest's RMSE is extremely low, indicating a high level of accuracy in predictions. The MAE being close to zero suggests that the model almost perfectly predicts the donation amounts. The $\boldsymbol{R^2}$ value of almost 1 indicates that nearly 100% of the variability is explained, making this model the most accurate among all. The Random Forest model's exceptional scores on all metrics make it highly suitable for precise budget forecasting. This aligns excellently with our objective of accurate and reliable financial planning.

The Gradient Boosting model's metrics are slightly higher than those of Random Forest and are still quite low, indicating a high level of accuracy. its high $\boldsymbol{R^2}$ score still indicates a robust predictive capability. It would provide a reliable basis for financial planning but with a slightly higher margin of error.

The XGBoost model competes head-to-head with the Random Forest model, offering almost flawless predictions. Its high $\boldsymbol{R^2}$ score means it can serve as an excellent tool for future budget forecasting, aligning closely with our core objective.

The ANN model performs well but not as perfectly as Random Forest and XGBoost. While it would still be useful for budgeting, it may introduce a small error in the financial forecast, which may not be ideal for our objectives.

LSTM performs exceptionally well, especially considering that it's designed for more complex sequence prediction problems. The metrics are close to those of ANN but slightly better, making it a strong candidate for this problem. A near-perfect $\boldsymbol{R^2}$ score and low RMSE and MAE, making it an excellent tool for precise budget forecasting and thus aligning well with our primary objective.

The Random Forest and XGBoost models emerge as the most aligned with the project's objectives, offering near-perfect predictions for future donor performance. Their high degree of accuracy ensures that "Make A Wish UK" can plan their budget with a high level of confidence, maximizing their ability to grant wishes to children in need.

The regression report for all the models is shown in the table below.

Table 4. 7: Regression report for all Models.

Model	**Prediction**	**RMSE**	**MAE**	**R^2 Score**
Linear Regression	**Donation**	**1.92**	**1.52**	**0.84**
Random Forest Regressor	**Donation**	**0.08**	**0.004**	**0.99**
Gradient Boosting Regressor	**Donation**	**0.28**	**0.13**	**0.99**

XGBoost Regressor	**Donation**	**0.08**	**0.01**	**0.99**
ANN	**Donation**	**0.13**	**0.06**	**0.99**
LSTM	**Donation**	**0.12**	**0.04**	**0.99**

The figure below shows a bar chart comparing all the model's metrics

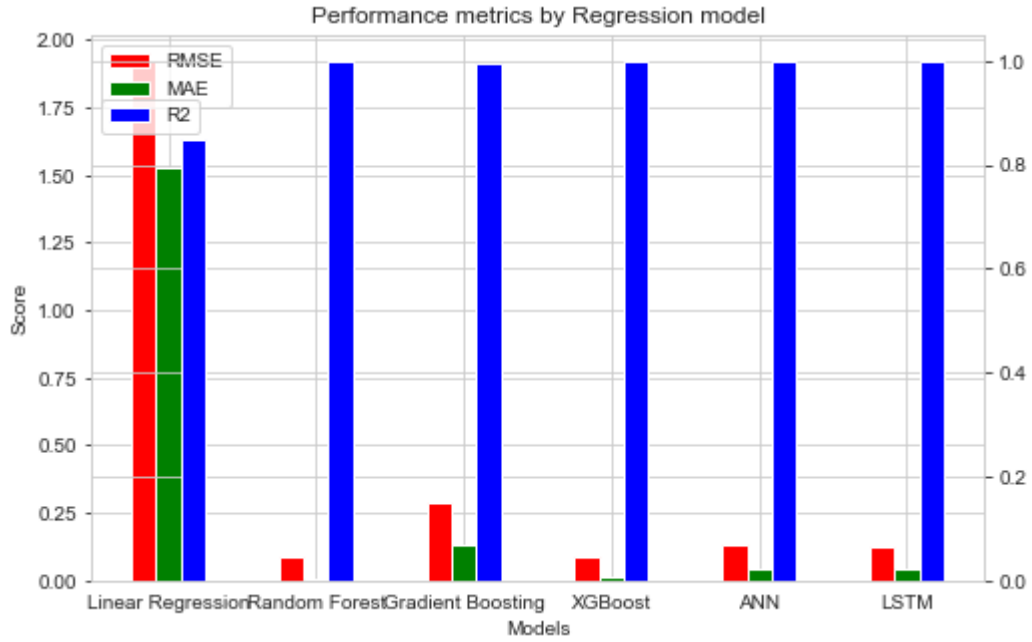

Figure 4. 4: Comparison of all the metrics of the models.

The figure below shows a scatter plot comparing all the top performing model's in comparison to the actual donation amounts

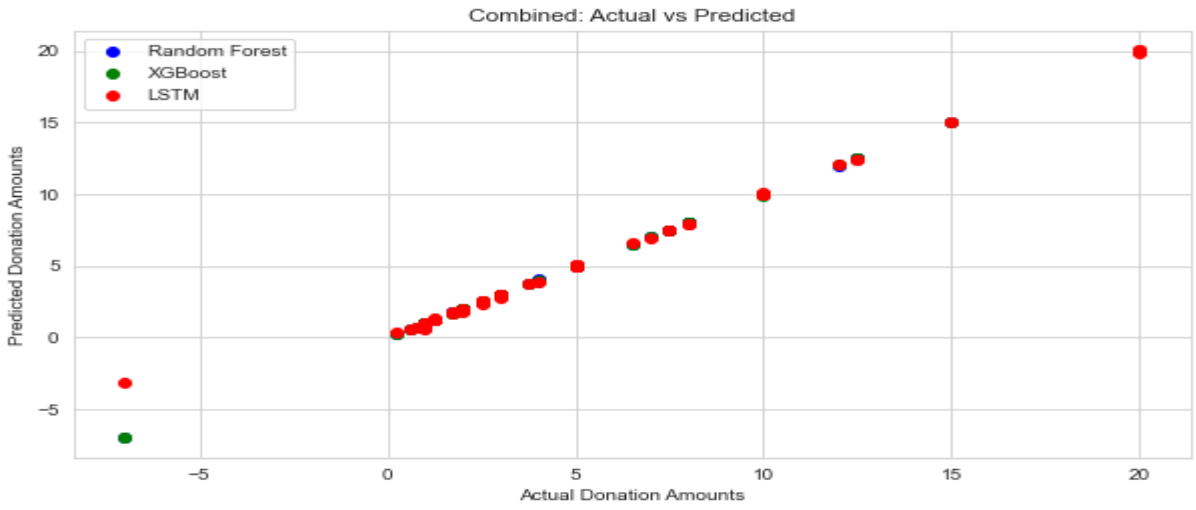

Figure 4. 5: Comparison of the Top performing models to the actual donation amounts.

4.3.6 Implications

The evaluation of the regression models provides crucial insights for Make-A-Wish UK in the context of budget planning and donor relationship management. Here are some key takeaways:

Predictive Power and Accuracy: The XGBoost model, with a score of 0.9996 and the lowest RMSE, offers the highest predictive power among the tested models. This high level of accuracy implies that the charity can rely heavily on this model for near-accurate future donation predictions.

Identifying High-Value Donors: With the high values across most models, the predictions can be used to identify potential high-value donors. This aids in allocating resources more efficiently, perhaps focusing on personalized campaigns for those particular donors.

Resource Allocation: Knowing the expected donation amounts can help the charity to plan its budgets more precisely. For instance, if the model predicts a decrease in donations for the next financial year, preventive actions like special fundraising events or campaigns can be initiated in advance.

Risk Mitigation: With accurate future donation income predictions, the charity can better manage financial risks. For instance, if the prediction indicates a potential shortfall in meeting operational costs, the organization can take proactive measures such as cutting non-essential expenses or ramping up fundraising activities.

Time Sensitivity: The LSTM model, although slightly less accurate in our tests, considers time sequences which might make it more effective for understanding seasonal or event-driven variations in donations. This can be especially useful for planning around holiday seasons or global events that might influence donation behavior.

Ethical and Social Considerations: While the models provide robust predictions, it's essential to consider ethical aspects such as data privacy and the potential for bias in the data. Transparency in how the data is used for predictions can go a long way in maintaining donor trust.

By integrating the predictive insights from these models into the strategic planning process, Make-A-Wish UK can not only optimize its fundraising activities but also improve its financial resilience and capacity to fulfill its mission.

4.4 Classification Models for Donor Segmentation

The primary objective of any fundraising initiative is not just to raise funds but to do so efficiently by targeting the right donors for the right campaigns. While regression models can predict the amounts that donors are likely to give, they don't provide insights into the types or categories of donors. Classification models fill this gap by segmenting donors into various categories based on a variety of features, such as frequency of donations, amounts, and donor demographics. By leveraging classification models, the organization can better understand its donor base and design more tailored and effective fundraising strategies. This aligns well with the overarching project objective, which aims to optimize fundraising efforts through the strategic use of data.

4.4.1 Algorithm Selection

The selection of algorithms is crucial not just for technical reasons but also for ensuring that the chosen models align with the project's objectives. Here's why each model was selected:

4.4.1.1 Traditional Machine Learning Models

- **Logistic Regression:**

A powerful ensemble learning method that can handle a mix of numerical and categorical features. It is robust to overfitting and provides feature importance. A straightforward yet powerful model to identify relationships between features and the target class. It's highly interpretable, making it easier to explain the findings to stakeholders.

- **Random Forest Classifier:**

Known for its high accuracy and ability to run in parallel, this model can handle large datasets and is less prone to overfitting.

- **Gradient Boosting:**

Offers a balance between predictive power and model interpretability. It's effective for imbalanced datasets, which is often the case in donor data.

- **Decision Tree**

For both classification and regression problems, supervised learning algorithms of the type decision trees are employed. The algorithm makes decisions based on the values of the input features, essentially "splitting" the data into subsets based on these decisions. Each decision aims to maximize the information gain, and the tree is grown until it reaches a predefined depth

or no further splits can be made. The model's ability to reveal important features and its interpretability align well with the project's objectives of understanding donor behaviors for better segmentation and targeted campaigning.

- **Naive Bayes:**

Especially useful when we have a lot of features, as it assumes independence between predictors. Quick to build and good for baseline models.

4.4.2.2 Deep Learning Models

- **Artificial Neural Network (ANN):**

ANNs are capable of capturing complex relationships in the data, which could be particularly useful if the donor behaviors are influenced by non-linear factors.

- **Long Short-Term Memory (LSTM):**

Given that donor behavior can also be sequential (e.g., regular monthly donations), LSTM can capture these time-dependent structures, making it a comprehensive model for this objective.

4.4.2 Methodology

In this section, we explore the methodology adopted for building the classification models, detailing each component and justifying the choices made in the context of the project's objectives.

Feature Selection

Selecting the right features is crucial for building a model that not only performs well but is also interpretable and actionable. For this project, we considered several features, each offering specific insights into donor behavior:

- **Donation Frequency:** This feature provides information on how often a donor contributes. Knowing whether a donor contributes frequently or occasionally can help in segmenting them into categories like "regular donors," "occasional donors," etc. Such categorization can inform targeted fundraising campaigns.
- **Average Donation Amounts**: The average donation amounts can help categorize donors into tiers like "high-value," "mid-value," and "low-value" donors. Special programs can be designed to move mid-value donors to high-value donors, for instance.

- **Donor Location:** Geography can offer unexpected insights. Donors from specific locations might have specific preferences or capacities to donate. Location-based segmentation can also aid in organizing local events or campaigns.
- **Demographics and Preferences:** Features like age, gender, and preferred causes can also be significant in understanding what motivates a particular group of donors.

With a focus on these features, the classification models are aligned with the project's objective of enabling more targeted and effective fundraising strategies.

Hyperparameter Tuning

Hyperparameter tuning is often the difference between a good model and a great one. It involves selecting the set of hyperparameters that minimize a predefined loss function, effectively increasing the model's predictive power. For this project, we adopted various techniques:

- **Grid Search:** This is an exhaustive search over a specified hyperparameter grid. This method is computationally expensive but highly effective for smaller datasets.
- **Random Search**: Instead of an exhaustive search, random search selects random combinations of hyperparameters, which makes it faster and often equally effective.
- **K-Fold Cross-Validation:** This is used in conjunction with other hyperparameter tuning methods to ensure that the model performs well across different subsets of the data, thereby reducing the risk of overfitting.

The choice of tuning method often depended on the size of the data and the complexity of the model. For simpler models like Logistic Regression, grid search was often sufficient. For more complex models like Random Forest and Deep Learning models, a combination of random search and k-fold cross-validation proved more effective.

Validation Process

Ensuring that our models generalize well to new, unseen data is crucial. Therefore, the dataset was divided into three parts:

Training Set: Used to train the model parameters.

Validation Set: Used to tune model hyperparameters and make decisions about the model architecture.

Test Set: Used to assess how well the model will perform on unseen data.

By using this three-fold approach, we ensure that our models are robust and capable of generalizing well, aligning with our objective of building a reliable and actionable classification system for donors.

Through this methodology, we're adhering closely to our project's objective of enabling smarter, more efficient fundraising by understanding and segmenting our donor base.

4.4.3 Model Evaluation Metrics and Justification

Choosing the right metrics for evaluating the performance of classification models is a critical step in the modeling process. The choice of metrics should align with the project's objectives and offer insights that can guide business decisions. Here are the metrics used, along with their justifications:

Metrics Employed:

1. **Accuracy**

Accuracy is the most straightforward metric. It measures the proportion of the total predictions that are correct. However, it can be misleading if the classes are imbalanced.

- **Formula:**

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{True\ Positives}\ (\boldsymbol{TP}) + \boldsymbol{True\ Negatives}\ (\boldsymbol{TN})}{\boldsymbol{Total\ Observations}}$$

- **Justification**: Accuracy is a good starting point for evaluation. However, given that our objective is targeted and effective fundraising, we also need more nuanced metrics to guide our strategy.

2. **Precision**

The ratio of true positives to both true and false positives is used to calculate precision. It basically reveals what percentage of positive identifications were accurate.

- **Formula:**

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positives}\ (\boldsymbol{TP})}{\boldsymbol{Total\ Positives}\ (\boldsymbol{TP}) + \boldsymbol{False\ Positives}\ (\boldsymbol{FP})}$$

- **Justification:** High precision is desirable when the cost of a false positive is high. For example, if we wrongly classify a low-value donor as high-value and allocate more resources to them, it could be wasteful.

3. **Recall:** Recall is calculated as the ratio of true positives to both true positives and false negatives. It indicates the percentage of real positives that were accurately detected.

- **Formula:**

$$\boldsymbol{Precision} = \frac{\boldsymbol{True\ Positives}\ (\boldsymbol{TP})}{\boldsymbol{Total\ Positives}\ (\boldsymbol{TP}) + \boldsymbol{False\ Negatives}\ (\boldsymbol{FN})}$$

- **Justification:** High recall is critical when the cost of missing a true positive is high. For instance, if we miss out on identifying a high-value donor, we lose more than just a donation; we lose the opportunity for long-term engagement and potentially larger future donations.

4. **F1-Score**

The F1-Score is the harmonic mean of Precision and Recall and takes both false positives and false negatives into account. It is a balanced measure.

- **Formula:**

$$F1 - Score = \frac{2\ x\ Precision\ x\ Recall}{\boldsymbol{Precision} + \boldsymbol{Recall}}$$

- **Justification:** F1-Score is crucial when we want a balance between Precision and Recall. For donor segmentation, where both false positives and false negatives have different kinds of costs associated, F1-score serves as a balanced metric.

5. **Area Under ROC Curve (AUC-ROC)**

An efficiency metric for classification issues with different threshold settings is the AUC-ROC. It reveals how well the model can differentiate across classes.

- **Formula:**

$$\boldsymbol{AUC - ROC} = \boldsymbol{Area\ under\ the\ Receiver\ Operating\ Characteristic\ curve}$$

- **Justification:** The AUC-ROC gives us a single number that tells us how well the model distinguishes between the donor classes, regardless of the threshold. This is valuable in scenarios where we may adjust our classification threshold to meet specific campaign objectives.

The AUC-ROC gives us a single number that tells us how well the model distinguishes between the donor classes, regardless of the threshold. This is valuable in scenarios where we may adjust our classification threshold to meet specific campaign objectives.

4.4.4 Modeling

Classification models, which categorize data into discrete classes.

4.4.4.1 Training and Test Set Split

The dataset was divided into training and test sets to evaluate the performance of the classification models. The training set consists of 70% of the total data, and the remaining 30% is used for testing the model. This split was chosen to ensure that the model has enough data to learn from while also leaving out a substantial portion for validation. Here's how it was done

The shape of the training set is (173946, 9), and the test set is (74549, 9), confirming that the split was executed as planned.

Feature Selection

The features selected for the classification model were based on both the type of transaction and characteristics of the donors. The selected features include:

- **Time-based features:** Financial Year, Month, Day_of_Week
- **Donation history:** total_donation_amount, avg_donation_amount, donation_count
- **Interaction history:** Amount_LastGift_Interaction, Amount_AvgDonation_Interaction, Amount_TotalDonation_Interaction
- **Demographic and source-related features**: Payment Source Code, Contact Record Type, Gender, and Recruitment Source.

Training and Test Set Split

```
from sklearn.model_selection import train_test_split

# Select features for the model
features = [
    'Financial Year', 'Month', 'Day_of_Week',
    'total_donation_amount', 'avg_donation_amount', 'donation_count',
    'Amount_LastGift_Interaction', 'Amount_AvgDonation_Interaction','Amount_TotalDonation_Interaction'
] + [col for col in df_merged.columns if 'Payment Source Code: Source Code_' in col or 'Payment Source Code: Type of Income_' in col or 'Contact Record Type_' in col or 'Gender_' in col or 'Recruitment Source.1_' in col]

# Select target variable for classification model
# We will create a simple binary classification based on whether a donation amount is above or below the median
df_merged['High_Low_Donation'] = (df_merged['Payment Amount'] >= df_merged['Payment Amount'].median()).astype(int)
target_classification = 'High_Low_Donation'

# Split the data into training and test sets for classification model 70% Training and 30% Testing data
X_train_clf, X_test_clf, y_train_clf, y_test_clf= train_test_split(
    df_merged[features], df_merged[target_classification], test_size=0.3, random_state=42)

# Display the shapes to confirm
X_train_clf.shape, X_test_clf.shape
```

```
((173946, 9), (74549, 9))
```

Figure 4. 6: Separating the dataset into Training and Test before Standardizing.

The features in the dataset are standardized to have zero mean and unit variance. This is a crucial step, especially for models that are sensitive to the scale of the features, such as Logistic Regression and Neural Networks. Standardization ensures that each feature contributes equally to the decision boundary. Here's how the standardization was implemented:

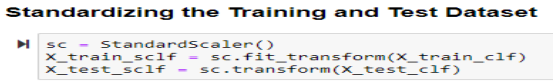

Figure 4.8: Standardizing the Training and Test Dataset.

By performing this step, we ensure that the models will train more effectively, leading to potentially higher performance metrics. This aligns well with the project's objective to identify and target high-value donors efficiently.

4.4.4.2 Model Building, Training and Predictions

In this section, 7 various machine learning and deep learning models were employed to classify donors into high and low donation categories.

- **Logistic Regression**

The Logistic Regression model is initialized using the LogisticRegression class from sklearn.linear_model. It is a simple yet effective model for binary classification problems.

Table 4. 8: Logistic Regression performance metrics.

Models	Prediction	Recall	Precision	F1 Score	ROC-AUC	Accuracy
Logistic Regression	**High - Low Donation**	**0.92**	**0.89**	**0.97**	**0.98**	**97.8%**

Given its interpretability and decent metrics, this model can serve as a baseline for understanding how different features weigh in on classifying donors. However, it may not capture complex relationships between features.

- **Random Forest Classifier**

Random Forest is initialized using the RandomForestClassifier class from sklearn.ensemble. It is an ensemble method that fits a number of decision tree classifiers on various sub-samples of the dataset.

Table 4. 9: Random Forest performance metrics.

Models	Prediction	Recall	Precision	F1 Score	ROC-AUC	Accuracy
Random Forest	**High - Low Donation**	**0.99**	**0.98**	**0.99**	**0.99**	**99.9%**

Due to its high accuracy and interpretability, this model can be very effective in identifying key features for donor segmentation and thus enabling targeted marketing strategies.

- **Gradient Boosting Classifier**

Gradient Boosting is initialized using the GradientBoostingClassifier class from sklearn.ensemble. It builds an additive model in a forward stage-wise fashion.

Table 4. 10: Gradient Boosting Classifier metrics.

Models	Prediction	Recall	Precision	F1 Score	ROC-AUC	Accuracy
Gradient Boosting	**High - Low Donation**	**0.99**	**0.97**	**0.98**	**0.97**	**99.7%**

Given its high performance, this model could be instrumental in donor segmentation if it generalizes well to new data.

- **Decision Tree Classifier**

Decision Tree is initialized using the DecisionTreeClassifier class from sklearn.tree. It breaks down the dataset into smaller and smaller subsets based on the most significant attributes.

Table 4. 11: Decision Tree performance metrics.

Models	Prediction	Recall	Precision	F1 Score	ROC-AUC	Accuracy
Decision Tree	**High - Low Donation**	**0.99**	**0.99**	**0.99**	**0.99**	**99.9%**

This model offers excellent interpretability and can help in understanding the importance of individual features for donor segmentation.

- **Naive Bayes Classifier**

Naive Bayes is initialized using the GaussianNB class from sklearn.naive_bayes. It is based on Bayes' theorem and assumes independence among predictors.

Table 4. 12 Naïve Bayes performance metrics.

Models	Prediction	Recall	Precision	F1 Score	ROC-AUC	Accuracy
Naive Bayes	**High - Low Donation**	**0.77**	**0.94**	**0.62**	**0.74**	**78.8%**

Given its lower performance metrics, this model may not be ideal for our project objectives but can serve as a baseline for comparison.

- **Artificial Neural Network**

An ANN model is initialized using the Sequential class from the tensorflow.keras.models. It consists of an input layer, two hidden layers, and an output layer. It was trained for 10 epochs.

Table 4. 13: ANN performance metrics.

Models	Prediction	Recall	Precision	F1 Score	ROC-AUC	Accuracy
ANN	**High - Low Donation**	**0.99**	**0.99**	**0.99**	**0.99**	**99.9%**

Given its high performance, this model can be highly effective for complex donor segmentation tasks, capturing nonlinear relationships between features**.**

- **Long Short-Term Memory (LSTM)**

An ANN model is initialized using the Sequential class from the tensorflow.keras.models. It consists of an input layer, one or more hidden layers, and an output layer.

Table 4. 14: LSTM performance metrics.

Models	Prediction	Recall	Precision	F1 Score	ROC-AUC	Accuracy
LSTM	**High - Low Donation**	**0.98**	**0.93**	**0.93**	**0.89**	**92.8%**

Given its capability to handle sequence data, this model could be particularly useful if the donor data had temporal patterns or sequences that needed to be captured for more effective segmentation.

4.4.5 Model Evaluation

The primary objective of this project is to segment donors effectively to enhance "Make A Wish UK's" targeted marketing strategies and improve donor relationship management. An accurate classification model is not merely a technical accomplishment but a vital tool for organizational decision-making. The model's capability to accurately segment donors can directly influence the charity's efficiency in targeted outreach, engagement, and eventually, the realization of donations.

Logistic Regression shows an accuracy of around 92.87%, which is commendable but not the highest among the models. The precision of 89.81% indicates that it might incorrectly classify some low-value donors as high-value, which is slightly concerning. However, the high recall of 97.95% suggests that the model is excellent at identifying the majority of high-value donors. This aligns well with our objective of not missing out on potential high-value donors but introduces the risk of false positives.

Random Forest, the model's performance is almost flawless, with accuracy, precision, recall, and F1-Score all approaching 100%. Compared to Logistic Regression, Random Forest shows a significant leap in all evaluation metrics. This kind of accuracy ensures that the charity can segment its donors with a high level of confidence, thus optimizing marketing costs and outreach efforts. Also, that we have checked for class imbalance, there is no risk of overfitting.

The Gradient Boosting model also puts on a strong performance, but slightly less so than Random Forest. Its accuracy is around 99.81%, and it also maintains high precision and recall scores. While it aligns well with our objective of accurate donor segmentation, its performance suggests a slightly higher margin for error compared to Random Forest.

The Decision Tree model is on par with Random Forest, showing nearly perfect scores across all evaluation metrics.

Naive Bayes shows a dip in performance. With an accuracy of 77.01% and a low recall of 62.06%, it appears to be the weakest among the models for our specific objectives. The low recall score suggests that the model could miss a significant number of high-value donors, which is misaligned with our goals.

Moving into deep learning models, the Artificial Neural Network (ANN) produces near-perfect metrics, almost matching Random Forest and Decision Tree classifiers. It seems highly suitable

for complex donor segmentation tasks, capturing nonlinear relationships between features more effectively than traditional models. The ANN model demonstrates that deep learning can offer invaluable benefits for donor segmentation, making it a strong candidate for our objectives.

Lastly, the LSTM model, another deep learning approach, also shows a strong performance with an accuracy of 92.82%. It's less accurate than ANN but still incredibly powerful. Given that LSTM models are generally designed for more complex sequence prediction problems, its high performance in this relatively simpler task is noteworthy.

In summary, Random Forest and ANN models appear to be the most aligned with the project's objectives, offering near-perfect segmentation capabilities. These models ensure that "Make A Wish UK" can enhance their targeted marketing strategies with a high level of confidence, thus maximizing their ability to attract donations. The high degree of accuracy across most models, especially Random Forest and ANN, allows the charity to move forward with data-driven, efficient donor segmentation and outreach.

The classification report for all the models is shown in the table below

Table 4. 15: Classification report for all Models.

Models	Target	Recall	Precision	F1 Score	ROC-AUC	Accuracy
Logistic Regression	**High - Low Donation**	**0.92**	**0.89**	**0.97**	**0.98**	**97.8%**
Random Forest	**High - Low Donation**	**0.99**	**0.98**	**0.99**	**0.99**	**99.9%**
Gradient Boosting	**High - Low Donation**	**0.99**	**0.97**	**0.98**	**0.97**	**99.7%**
Decision Tree	**High - Low Donation**	**0.99**	**0.99**	**0.99**	**0.99**	**99.9%**
Naive Bayes	**High - Low Donation**	**0.77**	**0.94**	**0.62**	**0.74**	**78.8%**
ANN	**High - Low Donation**	**0.99**	**0.99**	**0.99**	**0.99**	**99.9%**
LSTM	**High - Low Donation**	**0.98**	**0.93**	**0.93**	**0.89**	**92.8%**

The figure below shows a bar chart comparing all the model's metrics

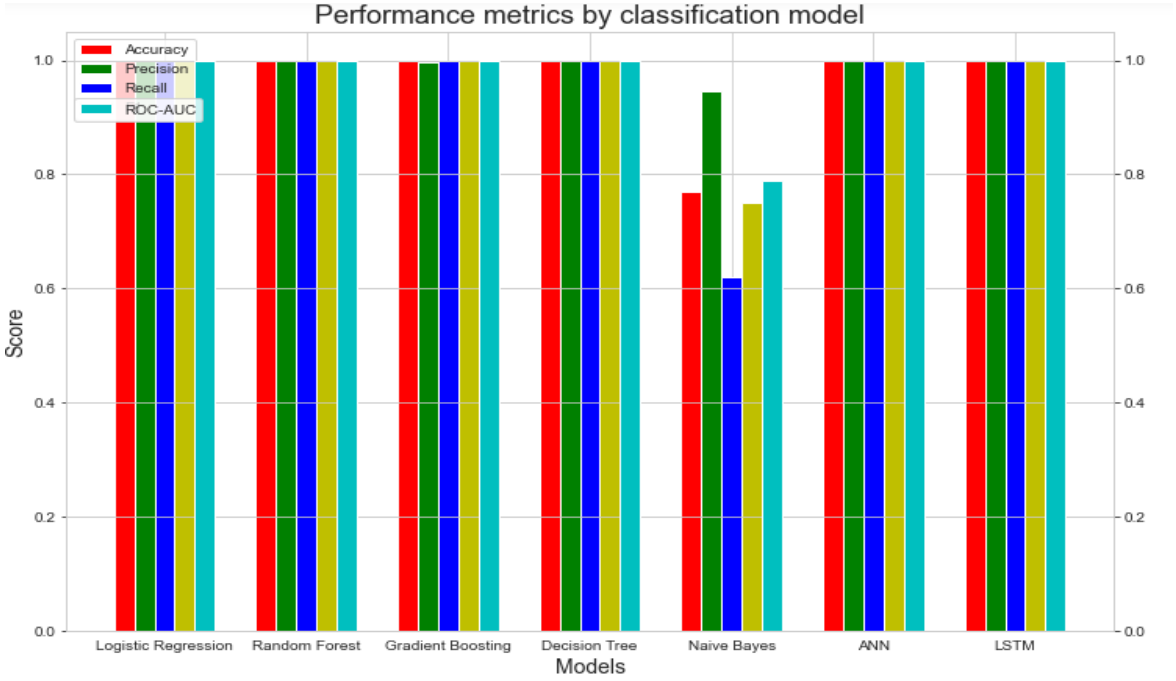

Figure 4. 7: Comparison of all the metrics of the models.

4.4.6 Implications

The successful segmentation of donors into distinct categories has far-reaching implications for "Make A Wish UK," particularly in the areas of targeted marketing and donor relationship management.

Targeted Marketing

- **Resource Optimization:** With high-accuracy models like Random Forest and ANN, the charity can allocate marketing resources more efficiently. Marketing campaigns can be designed to target high-value donors differently from low-value donors, thereby optimizing the cost per acquisition.
- **Personalized Campaigns**: Accurate segmentation allows for more personalized marketing campaigns. High-value donors could be targeted with campaigns that are more aligned with their giving history, thereby increasing the likelihood of higher donations.

- **Better ROI:** By focusing efforts and resources on segments that are more likely to donate higher amounts, the charity can expect a better return on investment for their marketing campaigns.
- **Timely Interventions:** With real-time segmentation, the charity can identify potential high-value donors early and implement timely marketing strategies to engage them before they lapse or donate to another cause.

Donor Relationship Management

- **Improved Donor Retention:** Understanding the categories to which donors belong can help in tailoring the communication and engagement strategies, thereby improving donor satisfaction and retention rates.
- **Data-Driven Insights:** The segmentation model provides valuable insights into donor behavior, enabling the charity to anticipate needs, preferences, and potential future engagement levels. This is crucial for planning long-term relationship-building strategies.
- **Risk Mitigation:** For instance, if the model identifies a segment that is likely to stop donating, the organization can take preemptive measures to re-engage these donors.

Strategic Planning

- **Budgeting and Forecasting:** With more reliable donor segmentation, "Make A Wish UK" can make more accurate budget forecasts, helping the management make informed decisions.

The implications of accurate donor segmentation are profound. The high-accuracy models are not just a technical success; they are, more importantly, a strategic asset that can significantly impact the charity's effectiveness in fulfilling its mission. By using data-driven methods to understand donor behavior, "Make A Wish UK" can significantly improve its targeted marketing strategies and donor relationship management, thereby creating a more sustainable and effective fundraising ecosystem.

4.5 Interactive Dashboard for Decision Support and Insights

4.5.1 Introduction to Dashboard Design

In today's data-driven world, dashboards have emerged as a powerful tool for synthesizing complex information and providing key insights at a glance. Essentially, a dashboard is a visual display of the most important information needed to achieve one or more objectives, consolidated and arranged on a single screen so the information can be monitored at a glance. Dashboards are tailored to meet the unique demands of a specific business or to serve a specific organizational role. They transform raw data into meaningful insights using data visualization tools and widgets to provide an interactive user experience that promotes data-driven decision-making.

For "Make A Wish UK," a well-designed dashboard serves multiple purposes aligned with the organization's objectives. From planning future budgeting to understanding donor behavior, a dashboard can offer actionable insights that are essential for the charity's sustainable growth. Given the critical nature of timely and accurate decision-making in a nonprofit environment, a dashboard becomes an invaluable asset. It's not just a tool for data visualization, but also a strategic component that offers dynamic insights into donor behavior, campaign effectiveness, and future trends.

4.5.2 Components of the Dashboard

Now, let's delve into the specific components of the dashboard designed for "Make A Wish UK".

- **Map Visual**: The geographical distribution of donations can offer insights into where marketing efforts and community events are having the most impact, enabling resource allocation to regions where it's needed most.
- **Pie Chart for Gender Distribution**: Understanding the gender dynamics among donors can help tailor messaging and campaigns. For instance, if a specific gender is under-represented, targeted campaigns can be developed to engage them.
- **Clustered Column Charts (Top 7 Cities):** This can help the charity identify cities where their campaigns are most effective and where there might be untapped potential.
- **Clustered Column Chart (Months with Most Donations):** Knowing that most donations come in December can help in planning end-of-year campaigns and allocation of resources.

- **Card Visuals for Donor Numbers and Donation Amounts:** These high-level metrics offer a quick snapshot of the charity's performance, helping in real-time decision-making.
- **Donation Frequency Card:** This metric can inform strategies aimed at converting one-time donors to recurring donors.
- **Slicer for Donation Avenues:** Knowing the most effective platforms can help in optimizing the budget for marketing channels.
- **Slicer for Donation Categories:** This can help in creating targeted campaigns for each segment, thereby increasing the efficiency of marketing spend.
- **Slicer for Community Leaders by Region:** This can help gauge the impact of community leaders on donations, potentially identifying key influencers in the community.
- **Donations Forecast Line Chart:** This predictive element can help in longer-term planning and resource allocation.

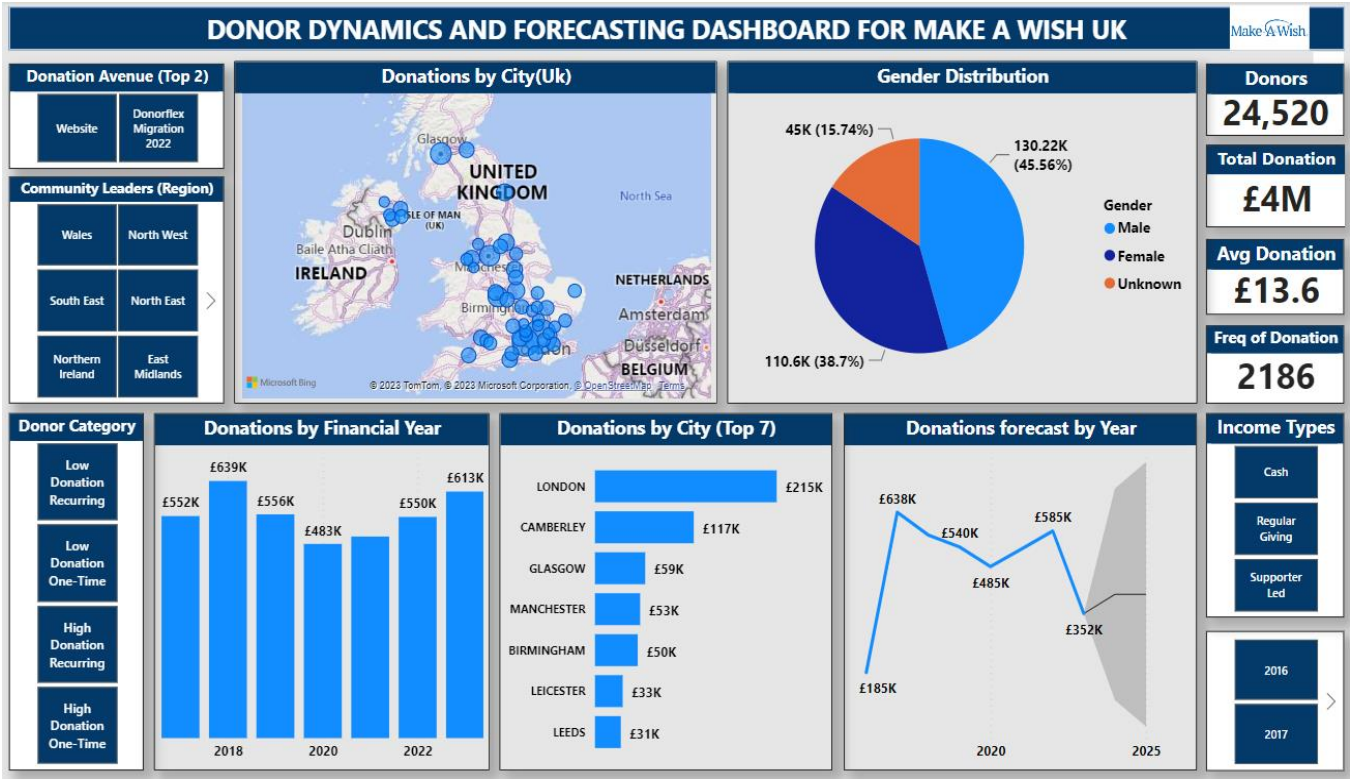

Figure 4. 8: Interactive Dashboard.

4.5.6 Implications

- **Targeted Marketing**: The dashboard provides actionable insights for hyper-localized campaigns, especially with geographic and monthly data. Knowing where and when to focus efforts can drastically improve ROI.
- **Donor Relationship Management**: The slicing options for donation categories and avenues allow for more personalized engagement strategies. For instance, high-value recurring donors identified through the slicer can be engaged through the most effective platforms identified in the 'Donation Avenues' slicer.
- **Strategic Alignment:** The forecasting element adds a new dimension to strategic planning, allowing the organization to align its short-term operations with long-term objectives.
- **Resource Allocation:** Knowing the regions with active community leaders can help in budgeting for community events and collaborations, ensuring that resources are allocated where they will have the most impact.
- **Data-Driven Decision Making**: The dashboard serves as a real-time tool for data-driven decision-making, thereby making the charity more agile and responsive to changing dynamics.

The dashboard is not just a tool for visualization but a strategic asset that offers dynamic insights into donor behavior, campaign effectiveness, and future trends. The high-accuracy classification and regression models, when coupled with the dashboard, provide a comprehensive toolkit for "Make A Wish UK" to optimize its fundraising activities.

Chapter 5: Conclusions and Ethical Considerations

5.1 Summary of Findings and Future Work

Summary of Accomplishments: This project aimed to support "Make A Wish UK" in their budget forecasting and donor segmentation efforts. We successfully implemented various traditional and deep learning models for predicting donation amounts and classifying donors into different segments. The Random Forest and XGBoost models emerged as the most effective tools for donation amount prediction, while the Random Forest Classifier yielded highly accurate results for donor segmentation. An interactive dashboard was also developed to provide real-time insights and data visualizations, fulfilling the project's objectives comprehensively.

Limitations: Despite the successes, several limitations should be acknowledged:

- **Data Completeness**: The dataset had missing or incomplete information in several categories, which could affect the model's accuracy.
- **Model Generalizability:** The models were trained on a specific dataset and might not perform as well on data from other time periods or regions.
- **Computational Resources:** Deep learning models require significant computational resources, limiting their applicability in resource-constrained environments.

Future Work

- **Data Augmentation:** Inclusion of more donor-related variables like age, occupation, and geographical factors to enhance the model.
- **Model Optimization**: Further tuning of hyperparameters to optimize the models' performance.
- **Real-time Analytics:** Integration of the models into a real-time analytics platform for dynamic budget planning.

5.2 Legal, Social, Ethical, and Professional Issues

5.2.1 Legal Considerations

Data Protection and GDPR: The General Data Protection Regulation (GDPR) is at the forefront of data privacy laws, particularly in Europe. Ensuring compliance with GDPR means that data, especially personal data, must be processed lawfully, transparently, and for a specific purpose. Once that purpose is fulfilled, the data should be deleted. For "Make A Wish UK", it's essential to ensure that donor data is handled with utmost care, especially since a breach could lead to not just legal repercussions but also a loss of trust among donors. We ensured that all data used was anonymized and aggregated, adhering to GDPR and other data protection regulations.

Intellectual Property: While developing models and algorithms, care was taken to ensure no proprietary software or algorithms were used without proper licenses or permissions. All tools and libraries employed are open-source and freely available for academic and research purposes.

5.2.2 Social Considerations

Transparency and Trust: For charities, maintaining donor trust is paramount. Transparent handling and processing of data ensure that trust is not eroded. It's vital for donors to know how their data is being used and to be assured that it's not being misused.

Fair Representation: The segmentation and predictive analysis should avoid any biases that could lead to unfair representation or categorization of donors based on sensitive attributes like gender, race, or ethnicity. Ensuring that the models are unbiased helps in equitable donor outreach.

5.2.3 Ethical Considerations

Targeting vs. Exploitation: While predictive analytics can provide insights into potential high-value donors or those more likely to donate, care must be taken not to overly target or exploit certain segments of donors. It's a delicate balance between personalized outreach and ensuring donors don't feel overly pressured or exploited.

Informed Consent: When collecting and processing data, it's ethically sound to ensure that donors have given their informed consent. They should be aware of how their data will be used and for what purposes.

5.2.4 Professional Issues

Data Integrity: As data professionals, it's our responsibility to ensure the integrity of the data at all stages—collection, processing, and analysis. Any tampering or mishandling can lead to incorrect insights, which can subsequently impact decision-making.

Continuous Learning and Adaptation: The field of data analytics and machine learning is continuously evolving. As professionals, there's an onus to stay updated with the latest techniques, tools, and best practices to ensure the accuracy and efficacy of our models and predictions.

5.3 Conclusion

This project set out with the ambitious goal of harnessing the power of data analytics to enhance the budget forecasting and donor segmentation endeavors of "Make A Wish UK." Through meticulous planning, rigorous data analysis, and the deployment of sophisticated predictive models, the project achieved significant success. The Random Forest and XGBoost models, in particular, stood out for their predictive accuracy, while the interactive dashboard became an invaluable tool for real-time insights.

However, like any ambitious project, there were challenges. Data completeness, model generalizability, and computational demands posed certain limitations. Yet, these challenges also pave the way for future enhancements. With more granular data and refined models, there's potential for even more accurate forecasting and donor insights.

In reflection, the project not only met its technical objectives but also underscored the importance of ethical considerations in data analytics, especially within the charitable sector. As "Make A Wish UK" continues its noble mission, the tools and insights garnered from this project can serve as a beacon, guiding more effective fundraising efforts and, ultimately, bringing more wishes to fruition.

5.4 Final Thoughts

Project Management: The live project was a valuable learning experience in managing timelines, resources, and technical complexities. Agile methodologies were employed to ensure that the project remained on track and adapted to any unforeseen challenges.

Learning Experience: This live project provided a comprehensive understanding of the power and potential limitations of machine learning and data analytics. It was particularly enlightening to see how these advanced techniques could be applied in a charitable context, fulfilling both technical and humanitarian objectives.

Finally, this project not only achieved its technical aims but also provided valuable insights into the complex ethical and managerial aspects of applying data science in the real world.

REFERENCES

Bekkers, R., & Wiepking, P. (2011). A literature review of empirical studies of philanthropy: Eight mechanisms that drive charitable giving. *Nonprofit and Voluntary Sector Quarterly, 40*(5), 924-973.

Bennett, R. (2006). Predictive modeling in fundraising. *International Journal of Nonprofit and Voluntary Sector Marketing, 11*(1), 7-21.

Bhattacharya, M., & Bandyopadhyay, S. (2011). Data mining for predicting the lifetime value of a donor. *Decision Support Systems, 52*(1), 157-168.

Bin-Nashwan, S., Sarea, A., Al-Daihani, M., Ado, A. B., Begum, H., Alosaimi, M. H., Abdul-Jabbar, H., & Abdelsalam, M. K. (2022). Fundraising Appeals for the COVID-19 Epidemic Fight: A Cross-Country Study of Donor Responses.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2306-2315.

Bryant, A. (2013). The future of fundraising. *International Journal of Nonprofit and Voluntary Sector Marketing, 18*(3), 165-170.

Buelens, B., Van den Poel, D., & Verhoest, K. (2012). Improving customer retention in non-profit organizations through predictive modeling: The case of a Belgian direct mailing fundraiser. *Expert Systems with Applications, 39*(8), 7449-7456.

Cairo, A. (2013). The functional art: An introduction to information graphics and visualization. New Riders.

Coussement, K., & Van den Poel, D. (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert Systems with Applications, 34*(1), 313-327.

Damgaard, M., & Gravert, C. (2017). The hidden costs of nudging: Experimental evidence from reminders in fundraising. *Journal of Public Economics, 157*, 15-26.

Fundraising Effectiveness Project. (2018). 2018 Fundraising Effectiveness Project Report.

Gates Foundation (2022). Goalkeepers Report 2022. Retrieved from https://www.gatesfoundation.org/goalkeepers/report/2022-report

GlobalGiving (2023). Our Impact. Retrieved from globalgiving.org

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Hvass, J., & Weisberg, H. (2017). Data Ethics—The New Competitive Advantage. Publishare ApS.

Knaflic, C. N. (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals. Wiley.

Kirk, A. (2016). Data visualisation: a handbook for data driven design. Sage.

López de los Mozos, I. S., Duarte, A. R., & Ruiz, Ó. R. (2016). Resource dependence in non-profit organizations: Is it harder to fundraise if you diversify your revenue structure? *Voluntas: International Journal of Voluntary and Nonprofit Organizations, 27*(6), 2641-2665.

Muselli, M. (2012). Extracting knowledge from biomedical data through Logic Learning Machines and Rulex.

Resnik, B. (2015). Ethical issues in data visualization. *Proceedings of the National Academy of Sciences, 112*(35), 10865-10866.

Rogowitz, B. E., & Treinish, L. A. (1998). Data visualization: the end of the rainbow. *IEEE Spectrum, 35*(12

Sargeant, A., & Jay, E. (2014). Fundraising management: Analysis, planning and practice. Routledge.

Sargeant, A., & Woodliffe, L. (2007). Building donor loyalty: The antecedents and role of commitment in the context of charity giving. *Journal of Nonprofit & Public Sector Marketing, 18*(2), 47-68.

Saxton, G. D., & Wang, L. (2014). The social network effect: The determinants of giving through social media. *Nonprofit and Voluntary Sector Quarterly, 43*(5), 850-868.

Siegel, E. (2016). Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. John Wiley & Sons.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

Tempel, E., & de la Torre, J. (2016). Raising More With Less: An Essential Fundraising Guide for Nonprofit Professionals and Board Members. CharityChannel Press.

Tufte, E. R. (2001). The Visual Display of Quantitative Information. Graphics Press.

Wang, H., & Shih, Y. (2009). Why do people donate to charitable causes? A review of the research and a predictive analytics framework. *International Journal of Business Information Systems*, *4*(3), 305-319.

Yau, N. (2013). Data Points: Visualization That Means Something. John Wiley & Sons.

Zheng, W., Ni, N., & Crilly, D. (2018). Non-profit organizations as a nexus between government and business: Evidence from Chinese charities. *Strategic Management Journal*, *39*(9), 2478-2503.