Comparative Analysis of Machine Learning and Sequential Deep learning Models in Higher Education Fundraising

by

Atsuko Umeki

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF APPLIED SCIENCE

in the Department of Electrical and Computer Engineering

Supervisory Committee

Dr. Alexandra Branzan Albu, Supervisor
(Department of Electrical & Computer Engineering)

Dr. Kin Fun LI, Departmental Member
(Department of Electrical & Computer Engineering)

ABSTRACT

Deep learning models have been used widely in various areas and applications of our everyday lives. They could also change the way non-profit organizations work and help optimize fundraising results. In this thesis, sequential models are applied in fundraising to compare their performance against the traditional machine learning model. Sequential model is a type of neural network that is specialized for processing sequential data. Although some research utilizing machine learning algorithms in fundraising context exists, it is based on the data extracted from the specific time window, which does not take time-dependency of features into account; therefore, time-series features are independent at each data point relative to others. This approach results in loss of time notion. In this thesis, we experiment with the application of time-dependent sequential models including Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU) and their variants in the fundraising domain to predict the alumni monetary contribution to the university. We also expand our study by including the architecture that treats time-invariant demographic data as a condition to the sequential layers. In this model, the time-dependent data is concatenated after running the sequential model. Sequential deep learning is empirically evaluated and compared against the traditional machine learning models. The results demonstrate the potential use of both traditional machine learning and sequential deep learning in the prediction of fundraising outcomes and offer non-profit organizations solutions to achieve their mission.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude my to supervisor, Dr. Alexandra Branzan Albu for guiding me in this work and her profound patience and understanding during my graduate studies. I am extremely grateful that she took me on as her student and continued supporting me throughout my study. Without her support I could not complete my thesis work studying while working full-time.

I would also like to thank my friends, Hadeer Ahmed and Maxine Gibson for their advice on my thesis work and their friendship when I felt overwhelmed and stressed with my study and regular work.

My special thanks also goes to our Director of Advancement Services, Stephanie Rowe and Associate Director, Gregory Churchill for their generous support in the use of their data and for their encouragement throughout my studies.

DEDICATION

I dedicate my thesis work to my sons, Eugene and Kent Umeki, who are my reason for living and love of my life. They supported and encouraged me during the challenging period of graduate study while working full-time. They never fail to brighten my life through the hardest time of my life. I also dedicate this work to my family in Japan, mother, Terumi Kano and sister, Takako Kimura for their unconditional love towards me. No matter how far away they live, they are always there for me at all times, in good times and not so good times.

Chapter 1

Introduction

The A.I. (Artificial Intelligence) in Advancement Advisory Council (AAAC) is an organization established in 2018 to help promote the evaluation, use, learnings, and outcomes from applying A.I. in fundraising [22]. According to the 2019 survey conducted by AAAC, 28 % of fundraising professionals are using or intend to use artificial intelligence in their work within the next six months. 42% do not have any plan to use A.I. within the year, and the remaining 30% have no plan to deploy or research A.I. at all. As these numbers indicate, the non-profit sector has not been a quick adapter of A.I. technologies. However, the survey also revealed that 89% of non-profit organizations agree that A.I. could make their work more efficient. The non-profit sector is behind the for-profit sector in the application of cutting-edge technologies, mainly due to their small budgets and inadequate staffing. The for-profit sector has invested vast resources in A.I. for personalizing, targeting, and marketing optimization. According to the 2019 CIO survey, Gartner reported that the number of enterprises employing A.I. has increased by 270% from 2015 to 2019, and 37 % of organizations have implemented A.I. in some form [20]. As in the for-profit sector, the non-profit organizations could benefit from A.I. application in their operation and enhance their Return on Investment (ROI). The lesser use of A.l. in the non-profit sector, in contrast to the for-profit sector, reflects fewer published studies that research the use of A.I. in the fundraising context.

Among these fewer published studies, one of the earlier ones is the research authored by Dietz L. H. (1985) [17]. Dietz used regression analysis to predict the alumni's financial contribution to Iowa State University. To our knowledge, his research paper was the oldest that applied machine learning in the fundraising context. The more recent studies in the fundraising domain include the research conducted

by Quin Chen (2010) [7] and Mark E. Walcott (2014) [48]. All these publications are based on the assumption that the fundraising data is time-invariant, meaning the information does not change over time and that charitable contribution in successive equal time spans is independent. The demographic information such as sex, marital status, education, relationship, and other types of data of similar nature is constant or rarely changes over the constituent's lifetime; therefore, this type of information is considered as time-invariant data. On the other hand, data such as giving amount, email open rate, click-through rate, event attendance are likely to change every year. Therefore, those types of information at each successive time steps are dynamic and subject to change. There is no research works that studies the use of time series data in fundraising.

A machine learning has been successfully used in many for-profit organizations and can also help non-profit organizations maximize revenue to advance their mission. A machine learning system that can increase fundraising productivity is desirable; however, the credibility of this kind is also questionable. Many factors determine its performance, including algorithm, data used, domain, features, and representations. We compare the performance among traditional time-invariant models and time-dependent models. We use the sequential model to incorporate a temporal sequence of events such as charitable contribution, open email count, click-through count, and event attendance. The comparative studies are conducted with Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and Recurrent Neural Network (RNN) variations as a sequential model. As a traditional model, we utilized traditional time-independent architecture similar to the classification model as presented by Diez [17], Chen [7] and Walcott [48].

1.1 Objective

Strategic marketing is a key component vital for any organization, and their business growth, both for non-profit as well as for-profit organizations. Non-profit organizations use similar marketing strategies to for-profit organizations to connect with their targets such as volunteers and donors. Many studies on marketing analytics are conducted to improve marketing results. In marketing analytics, machine learning is useful for discovering customer purchasing patterns. In fundraising analytics, the customer classification model in marketing domain is applied to analyze the donor's

giving patterns and optimize the efficacy of the target marketing as shown in Key's study [24]. Non-profit organizations can maximize the use of the limited fundraising resources and boost its return on investment (ROI) by studying prospective donors' giving behavior and predicting the future donations.

In our study, we employed the multi-class classification model as a traditional model where the amount of financial contributions are grouped into four classes: no donation, small donations, mid-size donations and large donations. The traditional model is applied to the time-invariant data collected at a point of time. Quin Chen [7] and Mark E. Walcott [48] used this type of model on the alumni data for fundraising purposes. They employed the traditional classification and linear regression models in predicting charitable giving patterns and amount of giving in the coming years. Their models utilize time-invariant demographic data and a set of donation data extracted from the original dataset for a specific time window. Their models assume that the giving amount is the same in each successive, equal-length of time period, meaning that the giving pattern and amount remain the same over time. Similarly, T. K. Das (2015) [15] structured his marketing classification model that classifies the customers, who will respond to the product offers in their model. The model assumes the shopping behavior is time-invariant in the same manner as structured in the models by Chen [3] and Walcott [4] in their fundraising analysis.

The traditional model has two major challenges: feature engineering and time-independent assumption. Firstly, while feature engineering is a vital component of successful traditional machine learning models, it is a laborious process and requires domain knowledge. Second, the model does not consider the time-variant relationship as discussed. Yang Jiang *et al.* [4] discuss that Deep Neural Networks (DNN) has an advantage over feature engineering for its capability to automatically detect essential features from the raw data. To handle the second time-independent challenge, there are two approaches. The first approach is the application of time-shifted windows. Lian Yam *et al.* (2017) [52] proposed the use of time windows to improve customer churn model performance. The customer churn model measures the immediate or future risk of customer cancellation. To predict future risk, they extracted training data from several time-shifted windows, and the final results are calculated by weighted average of time-shifted predictions. Their ensemble classification method outperformed the traditional time fixed model. The purpose of time-shifted windows is to predict non-stationary customer behavior. Their approach considers the time notion by incorporating several time-shifted windows, but each time window is still treated as

independent in the model. Another approach to account for time-dependency is the use of sequential model. To the best of our knowledge, there are no research works that utilize time dependent data in fundraising. In this study, we experiment the use of deep sequential neural network to handle the aforementioned feature engineering and time-Independent challenges experienced with the traditional model and examine its performance against the traditional model.

Recently, more researchers have studied the application of sequential model on time-variant data. Sequential models have various applications, including stock price prediction [46], speech recognition [36], machine translation [13], and sentiment classification [10] [28]. The sequential models are usually applied on the sequence of one type of feature, for example, historical stock prices, utterances, or customer comments. It becomes challenging when data consists of both static time-invariant and dynamic time-variant events. Liu *et al.* (2018) [44] and Wang *et al.* (2019) [8], for example, used sequential models, such as Long Short Term Memory (LSTM), on the historical data which contains a mixture of static and dynamic data. Liu *et al.* [44] experimented with real-world clinical data containing lab tests, diagnosis, drug administration, and patient demographic information for their death prediction. Their clinical data consists of both static demographic information and sequential clinical data such as blood test results. Wang *et al.* used LSTM for their study in medical crowdfunding. Their dataset also includes both static demographic data and time-varying crowdfunding data. The dynamic data contains how many times the case was viewed, how many unique donors donated every day, etc. They demonstrated that sequential models on the data mixed data are quite effective. Our study objectives are to examine the use of sequential learning in fundraising and compare its performance against traditional unimodal models. The dataset used for our study is the alumni records collected by the University of Victoria Advancement Services. The research conducted at the university conforms with the university's policies, and all forms of data are in completely anonymous formats (Protocol Number: 19-0025-02).

Our contributions include:

1. We explored an application of a sequential learning architecture in fundraising analysis. To our knowledge, it is the first attempt for fundraising domain.

2. We extend our study to experiment an architecture to handle both static and dynamic data by concatenating time-invariant demographic data followed by

temporal layers.

3. We critically compare the performance of sequential and traditional approach and examine the potential of sequential learning in the fundraising domain.

The rest of the thesis is organized as follows. Chapter 2 reviews the related literature on machine learning and sequential learning models in fundraising, marketing and crowdfunding. Chapter 3 provides a descriptive analysis of our dataset and discusses its limitations. Chapter 4 describes the considered sequential learning algorithms in detail, namely Recurrent Neural Network, Long Short Term Memory, and Gated Recurrent Unit and the sequential learning variations including bi-directional and time-distributed model. Chapter 5 discusses the model assessment methods and the study results. Chapter 6 summarizes our findings and insights.

.

Chapter 2

Literature Review

In this chapter, we review literature applying machine learning techniques in three areas. The first part 2.1 covers the related literature in fundraising analytics, including the commonly used classification model in the traditional time-invariant machine learning approach. The second part 2.2 discusses the literature using the sequential model in marketing, as well as discusses two different approaches to e-shopping analysis in target marketing. The last part 2.3 is about works that uses model architecture that combine time-variant and time-invariant data.

2.1 Marketing for a Non-Profit Organization

In 1975 Kotler [26] presented a strong case for the non-profit sector to introduce marketing in their organizations. Kotler firmly delivered that "Within another decade, marketing will be a major and accepted function." In 2003 Andreasen and Kotler [34] claimed that the need for non-profit organizations to generate more revenue to fulfill their mission drove them to adopt marketing principles and practices in their operation. However, the adaptation of marketing is still slow in many non-profit organizations. Shah and George [16] conducted an empirical study to evaluate the impact of marketing efforts on non-profits' performance outcome(s). Their findings are discussed in three areas, conceptual and practical contribution, economic implication of marketing, and public policy implication. In all areas, their findings strongly indicate the benefits of using marketing in non-profit organizations and suggest that non-profit leaders should invest in marketing to boost non-profit organizations' performance. Section 2.1 discusses the use of machine learning in fundraising, and 2.2

discusses the literature using sequential deep learning in marketing. We investigate the potential use of sequential learning for fundraising analysis by reviewing the existing research works.

2.2 Machine Learning in Fundraising

Direct mail and email marketing campaigns are widely used in fundraising marketing to connect to the target audience. The idea of target marketing is to send out the fundraising message to those who benefit from it or are likely to respond to such messages. It is a cost-effective way to optimize resource allocation. Machine learning helps make an educated guess about future activities such as customer purchase behavior or donor giving behavior. By adopting a machine learning approach, the chances of targeting the right group of people are enhanced. The traditional approach to fundraising analysis commonly uses classification and regression. The objective of classification model is to identify individuals who are likely to donate. This type of classification model uses a binary (donate, not donate) or multi-class classification. Another common approach is to estimate the amount of the donor's next financial contribution by using regression. Both methods are based on the donor's demographic information and past giving data. Traditional approaches in fundraising analysis include donor classification and response model that identifies responders to campaign messages.

Studies on fundraising classification are limited in numbers; however, more research studies about response model are found in direct marketing. The response model is a classification model that classifies the customers likely to respond to the next marketing campaign based on the customer information. Direct marketing is a popular fundraising marketing tool to connect with prospective donors. Key (2000) [19] utilized the response model in the fundraising context. The author proposed applying Probit regression and Bayesian models that sets the target variable based on the recent gift information. Probit regression has binary dependent variables and uses the cumulative distribution function of the standard normal distribution.

$$E(Y \mid X) = P(Y = 1 \mid X) = \phi(\beta_0 + \beta_1 X) \tag{2.1}$$

where

X = independent variables

Y = target variables

The target variable is the variable whose values or classes are predicted, while the predictor variable is the variable whose values are used to predict the target variable. In Key's model, the individuals who donated are coded as 1, while those who didn't donate are coded as 0. The predictor variables are an individual's demographic information (age, gender, marital status, etc.) combined with credit and census data. The predictor variables with thin correlations are removed by examining the correlation between predictor variables and the target variable. Multi-collinearity occurs when the predictor variables among themselves are highly correlated with each other, which causes unstable estimates. The author tested multi-collinearity among the predictor variables and removed correlated variables to avoid model instability.

The model structure Key [24] used is called response model, which is used to evaluate the contributing factors for prospects to respond to the next campaign appeals, and make financial contributions. The model was built on the datasets from the Catholic School and Metropolitan Museum in the United States. The paper does not cover descriptive statistics of the dataset, which usually reveals facts about data including data imbalance, missing data, skewness and anomalies. The fundraising datasets are often highly imbalanced with a higher proportion of non-responders compared to responders. The discussion of common issues associated with response model and ways to handle them is not addressed in the paper.

Walcott (2014) [48] took a similar approach. The research questions for his study are:

1. What factors are most likely to predict the likelihood of alumni making a financial contribution to their alma mater?

2. What factors are most significant in predicting the donation amount alumni will contribute to their alma mater?

Similarly, Walcott used the demographic data (age, gender, marital status, etc.), connectedness to the college, and alumni experience. In his study, he investigated factors that affect the likelihood of becoming an alumni donor using descriptive statistics followed by one-way ANOVA to test the strength of correlation between predictor variables and the target variable. The variables that have a significant correlation with the target variable were selected for further analysis. The author used binary Logistic Regression to select the best 14 predictor variables. To obtain the best subset of predictor, he used the method called 'best subset selection' by adding a variable until

it reaches the best accuracy. To answer the second study question, he ran Multivariate linear regression to evaluate the variables that best predict the first gift amount. After the model evaluation, nine predictor variables were chosen in the multivariate linear regression.

For the Logistic regression model, "1" is assigned to the alumni who have previously donated, while "0" is assigned to the alumni who have never donated. The dataset used for the analysis is highly imbalanced, with 71% non-donor records. The overall model accuracy is 75%, and the recall rate of donor and non-donor indicates the model correctly classified non-donor is 92.5% and for donor it is 30%. The argument Mark Walcott made on the validity of the model is weak because the dataset contains 71% non-donors, suggesting that even if the model predicts all the records to be "0", you still get 71% accuracy. Although the accuracy is relatively higher, the recall rate for the minority class (donors) is pretty low at 30%. However, in fundraising studies on highly imbalanced data, the low recall rate is not rare.

Chen (2010) [7] conducted a comparative analysis over four types of models, namely multiple regression, logistic regression, neural networks, and support vector machine (SVM), over their predictive performance. The study goal is the same as in Key [24] and Walcott [48]; first, to identify donors and non-donors and second, estimate the future giving amount. Chen's main contribution was to apply SVM in fundraising analysis.

Ha *et al.* [9] proposed the application of a bagging neural network model in the direct marketing response model. Response model is a popular tool for the fundraising campaign appeals. They argue that a complex neural network model has a large variance and a small bias while a simple model has a large bias and a small variance; this is called a 'bias-variance' dilemma where bias improves at the expense of the other and vice versa. Bias is the amount by which a model's predicted value differs from the target value, while variance means to what extent a randomly selected variable differs from its expected value. One way to mitigate this dilemma is to combine multiple networks. Therefore, they adapted a combined method that merges a neural network and bagging, called a bagging neural network. The bagging process precedes a neural network, which randomly samples N records with replacement L times. This process is called bootstrapping. Each bootstrap is used to train a neural network; therefore, there will be L neural networks. The final outcome is obtained by averaging a result from each neural network for regression or by majority voting for classification. In their study, the majority voting method was applied to select the final outcome.

Their contribution is to combine neural network and bagging to eliminate the time-consuming model selection because an ensemble of over-fitted neural networks tends to cancel out each other's peculiarities resulting from particular datasets. [9]

The majority of the early fundraising models use logistic regression, Linear regression, support vector machine, and neural networks. Those algorithms are also popular choices for the classification, response, and churn models in marketing. of those studies mentioned above attempted to consider the time notion in their models. As the usage of A.I. in e-commerce proliferates, the sequential deep-learning application has dramatically increased on time-series data from e-commerce marketing. The study on the deep learning approach in the non-profit sector is non-existent or has not advanced since the early days of fundraising analytics. The fundraising goal is to identify the patterns influencing donors' giving behavior over time. The application of sequential learning could shed light on fundraising marketing. The section 2.3 reviews how sequential learning is used in marketing.

2.3 Sequential Learning in Marketing

Time series data is a set of observations collected in even intervals over time, while cross-sectional data is collected at a single point in time. There may be a correlation between observations in time series data. There is an abundance of research works utilizing sequential deep learning models in for-profit marketing. Non-profit organizations use many of the same for-profit marketing strategies to connect with donors and volunteers. Fundraising data also contains time-series data. However, there are no research works utilizing sequential data to my knowledge. This section introduces some literature that studies the application of sequential learning to the time-series data from the e-commerce datasets. The literature that uses its application in marketing provides ideas how sequential learning can potentially be used in fundraising marketing.

In recent years, the drop in computational costs rekindled a hype in deep learning neural networks. Among various types of Deep Learning (DL) algorithms, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have marked an outstanding improvement in performance and have many examples of real-world applications. Neural Networks can train with different data types, including images, voices, texts, and objects. Moreover, some types of deep learning algorithms are capable of solving time-series data. Both marketing and fundraising data contain

two types of time-related data, dynamic and static time-series, specifically customer behavior in the context of marketing and donor's giving behavior in fundraising. The traditional classification model treats the features as static, including customer's purchasing history or donor's giving history over time. Recent studies in marketing incorporate these time-variant features in the model. Particularly, sequential neural networks are capable of capturing time dependencies. A sequential model is a model whose inputs or output have a sequential dependence, for instance, predicting the next purchase item from a sequence of previously purchased items. The following section discusses research papers that apply sequential learning to marketing.

Cui *et al.* (2018) [45] proposed a model that combines CNN and RNN for modeling the customer's online visit behaviors to predict the probabilities of conversion rates. The term 'conversion rate' describes the percentage to get customers to do what the marketers hope for them to do in marketing, for example, purchasing products, upgrading their services, etc. In fundraising, conversion rate means the percentage of a non-donor to donate or a donor to upgrade their monetary contribution. Cui *et al.* applied CNN to encode the keywords users employed when searching in Google search engine and the site names. CNN captures character-level features from each keyword in the search or URL name. The character-level features make it possible to utilize various types of words, abbreviations, or typos in the searched keywords. The embedded feature vectors extracted by CNN are fed into RNN to model the in-session customer behavior. Next, the Monte Carlo process is used to simulate the customer shopping journey to predict the conversion rate. The model can be used for the target advertising to optimize the advertising cost and resources.

Salehinejat *et al.* (2016) [39] proposed the customer lifetime value (CLV) model in which the customer shopping behavior is studied with RNN using the customer's recency, frequency, and monetary value (RFM) data. Because the time of first purchase differs among customers, they set the lower and upper limit of the time interval. The lower limit is the start point while the upper limit is the end point in their study. 'Recency' means how recently a customer made a purchase, 'frequency' is about how often a customer makes a purchase and 'money value' denotes how much money a customer spends for purchases. The model consists of one input layer, one hidden layer with ReLu activation, and the output layer. They used an autoencoder to get features from the customer loyalty number and pass them along with RFM data as a sequence at a time, t. The target of the Model is RFM values at the time, t+1. At each time step, customer lifetime value (CLV), R, F, M data for each customer go

into the input layer. The time interval can be weekly, bi-weekly, monthly, etc. They also set the lower and upper limits for the sequential time because the first purchase time differs among customers. They [39] claim that their proposed model is unique in its use of RFM and CLV data to feed into RNN. Both RFM and CLV data are also often used for fundraising analysis; thus their approach could be used to estimate donors' lifetime financial contribution in the fundraising domain.

Lang and Rettenmeier [38] demonstrated the advantage of using a sequential algorithm, RNN, to predict customer behavior in e-commerce shopping. The model was built on a large-scale e-commerce dataset, and the empirical results were compared to the traditional time-invariant model. They demonstrated the advantage of using a sequential model, RNN, over time-invariant models. The model was built on large-scale real-world datasets from European online fashion platform. Their focus is to predict the probability $p((o^u \mid x_1^u + x_2^u \ldots . x_T^u))$ of consumer u to place an order o^u where x_1, $x_2 \ldots . x_T$ denotes customer events at time step 1, 2, ... T. the customer events x have an action type such as product views, cart-additions, orders, etc., a timestamp and other session related information. For simplicity, they apply the RNN model on the product-specific orders. Sessions from six consecutive weeks in the spring of 2016 are used for the experiment. The sessional data from the first four weeks are used for the training, and the data from the remaining two weeks are used for the testing. They used RNN with a single LSTM layer and combined the binary time-invariant features at the last step of the time sequence.

Batmaz *et al.* [3], in their literature, provides a comparative review of existing recommender systems literature. The authors discuss that the popular collaborative filtering (CF) approach in recommender systems assumes that people who agree to their taste in the past would also agree in the future. As the CF assumption indicates, they claim that a user's current browsing history affects his/her future purchasing behavior. They also introduced another literature that captures this time-dependency by applying Recurrent Neural Network (RNN) and its family algorithms. Likewise, we believe that giving behavior is affected by past giving. To be more specific, alumni who are more engaged with the university (event attendance,email open, clickthough) in the past may donate in the future or alumni who donate a significant amount in this year may give less or no financial support in coming years. They claim that sequential learning can identify the sequential patterns better than the traditional models with cross-sectional data (non-time-series).

The literature mentioned above demonstrate that the sequential models offer a

legitimate advantage over the other methods in marketing.

2.4 Deep Learning in Crowdfunding

The success of deep learning practices reshaped real-world applications in many fields, including e-commerce and crowdfunding. Sequential deep learning has been successfully used in the crowdfunding sphere, among many other areas. This section discusses the literature that is the theoretical foundation of our study using a similar paradigm. Wang *et al.* [8] studied the application of sequential learning in crowdfunding. Crowdfunding is to raise small amounts of capital from a large number of people. It was first created for entrepreneurs to attract small-sized investments. Crowdfunding has been used to raise funds for a wide range of for-profit, entrepreneurial ventures, medical expenses, travel, and others [51]. The National Council of Nonprofits (Washington DC, USA) [32] projects that crowdfunding will become a $90-96 billion dollar industry by 2025 and will be a valuable tool for fundraising for charitable and nonprofit sectors. As the report indicates, crowdfunding is recently gaining popularity in the non-profit sector for its fundraising campaigns. Crowdfunding is rich in information about a project profile, including text, images, personal data, and metadata in multimodal form.

The goal of the model is to predict the fundraising outcome at the early stage of the crowdfunding campaign. The uncertainty over the fundraising outcome influences a fundraiser's decision on whether to receive medical treatment or search for other funding sources. The model aims to predict the outcome in a timely manner for the fundraiser to seek alternative financial means to receive treatment before the campaign's end date, which is determined at the launch of the crowdfunding projects. Their contribution is to combine time-invariant profile data and time-variant communication flow data, aiming to improve the model performance. The dataset includes the fundraiser's demographic information (age, gender, location) and insurance status as time-invariant data and the number of replies to donors, post updates, direct messages, post shares, and views from the daily flows on the crowdfunding site as time-variant data. Their proposed model architecture combines Convolutional Neural Network (CNN) for the time-invariant data and Recurrent Neural Network–Long Short Term Memory (RNN-LSTM) for the time-variant data. The authors review the advantages and disadvantages of four methods described below to combine time-

variant and time-invariant data in their study.

1. Treat the static features as dynamic at each stage.
2. Concatenate the static features with the output of the hidden state at the last time step.
3. Apply the static time-invariant features to initialize the first hidden layer (t = 1) and concatenated with the dynamic features at the following stages (t > 1)
4. Concatenate the static features with the hidden state at each time step.

The authors adopted the third method, the time-invariant features are fed into CNN one time at the beginning stage of the training process. The flattened profile data by CNN is combined into the RNN-LSTM layer. The authors achieved highly accurate prediction of the project success.

Zheng *et al.* performed a comparative analysis of several machine learning algorithms, namely, MLP, logistic regression, decision tree, random forest, SVM and KNN. The observation window for the study is from 2015-01-01 to 2018-11-29. The input and output data are treated as static. In their research work, the authors did not include dynamic attributes such as the founders' social media attributes (Twitter,Facebook or Flickr), dynamic project attributes (updates, comments), social promotion (number of followers); however, they acknowledge the necessity for a comprehensive evaluation of how various factors including dynamic attributes affect crowdfunding outcomes.

Cheng *et al.* [50] propose models that predict if crowdfunding projects result in success by evaluating data solely from the time-invariant project profiles. The authors designed the model on multimodal information from the project profiles such as texts, images, and metadata. Their focus is to evaluate the contribution of images to the predictability of the CNN model. Contrary to the work by Wang *et al.* [8]), the research work does not depend on the time-dependent data, i.e. time-variant daily flows on the crowdfunding site.

The definition of success is whether the project creator can reach their monetary goal by the end of the limited campaign period. They [50] use the pre-posting information to predict the campaign outcome before the project profile is even posted. The main reason for using pre-posting information is that the project creators will be able to know the likelihood of their success or failure and modify their profile

proactively or find alternative means to raise funds. The model's framework consists of three branches: encoding textual inputs, encoding image contents, and encoding metadata. Each branch consists of either Convolutional Neural Network (CNN) subnet or fully connected hidden layers. At the end of the branch stream, three feature maps are concatenated into one. The data was collected from the crowdfunding site "Kickstart" from 2015 to 2018. The data from 2015 to 2016 was used for training, 2017 for validation, and 2018 for testing. Despite the fact that the dataset consists of data from multiple years, they are treated as time-invariant data. The use of the time shifted data for training, validation and testing assures unbiased estimate of model accuracy. Their contribution is the use of multimodal information, which was a first attempt in time-invariant architectures. Nowadays, more images are used in not only crowdfunding projects, but also marketing and communication. Their empirical studies illustrate that the use of multimodal data including visual images could achieve superior performance. The fusion of two models, the sequential and multimodal data models, is another approach we could experiment with in fundraising. The multimodal information, such as text and images from the email communication, could be included in the model.

Srinivasan (2020) [42] explored successful crowdfunding based on three key factors: enticement, experience, and engagement. The author examined the effectiveness of content, its sentiments, reward, tangibility, funder belief, and founder-funder engagement by analyzing Kickstarter data. The author applied an ensemble deep-learning model and achieved 93% accuracy. Srinivasan's study shares similarities with the aforementioned literature in the study objectives to examine contributing factors in crowdfunding. The author extended the research and included an additional factor, i.e., how the tangible rewards incentivize funders.

Sequential leaning can be applied in various areas for supervised learning problems, containing a sequence of data as input or output. In this section we introduced the related studies using the sequential model in marketing and crowdfunding fields. These existing studies illustrate how to use sequential models by using data from the e-commerce online visit behavior or fundraisers interaction on the crowdfunding platforms to predict the sequence of the future behavior. Both for-profit and non-profit organizations share the same goal, generating revenue. Marketing strategy provides both for-profit and non-profit organizations an edge to achieve their goal. Optimization of marketing return on investment can be achieved by applying machine learning techniques. alumni data consists of time sequence inputs and outputs; our

study offers opportunities to explore a new approach for non-profit marketing.

Chapter 3

Methodology

Deep sequential learning has successfully been used in many areas including e-commerce and crowdfunding marketing strategy; however, there are not many studies on the application of deep sequential learning in the fundraising domain. This research study investigates how the introduction of sequential learning and combined data of time-series features and time-invariant demographic features influence the overall model performance compared to the traditional time-independent model in fundraising. This chapter gives an overview of the theoretical background of sequential learning and describes how our study was conducted.

3.1 Model Description

Two approaches were employed for a performance comparison, the traditional time-invariant approach and the sequential dynamic (time-variant) approach. Neural Network (NN) and Support Vector Machine (SVM) were used for the traditional time-invariant approach. NN and SVM have the ability to learn non-linear and complex relationships. For the time-variant sequential model, we applied three variations of Long Short Term Memory (LSTM) and two variations of Gated Recurrent Unit (GRU).

The goal of the traditional model in fundraising is commonly to classify constituent's response to campaign appeals in the binary (eg. donate or not donate) or multi-class model (eg. no donation, less than $1000, between $1000 and $10000, and greater than $10,000), or to predict future numerical donation amount by employing linear regression. Some information such as giving information and event partici-

pation change over time, but they were treated as time-invariant in the traditional models. As discussed in Chapter 2, Key [24], Chen [7], and Walcott [48] studied the relationship between demographic features and the target donation amount to predict multi-class donation amount and to estimate the numerical value of donation amount by using the time-invariant observations collected from the specific time window. In our study, we introduce time-series data in the model to identify giving patterns over time and consequently classify the range of their giving amounts including no donation in the coming year. To understand the sequential giving patterns, we used LSTM and GRU for the reason that LSTM and GRU are capable of learning and remembering over long sequences of inputs. Our first study objective is to examine the use of a sequential model in fundraising. The predicted amounts are grouped into four classes based on the 2020 donation amounts. Our study assumption is that time-series observations are correlated with one another, so each observation does not work independently. By introducing time-dependent data in the model, we examine if sequential learning is capable of capturing giving patterns over time.

Our second objective is to examine the performance of the model designed to handle both time-variant and time-invariant input in the models. The demographic features (age, sex, education, etc.) are not time series in nature, which are typically not included in the sequential deep learning architecture. Demographic data contains valuable information about donors. Our study question is whether adding static demographic data leverages the performance of sequential learning. To examine it, we concatenated time-series donation features and time-invariant demographic features in our models. As discussed in Chapter 2, Wang *et al.* [8] presented four existing and proposed approaches to combine time-variant data and time-invariant data. Their approach is to concatenate the time-variant information before running sequential layers. Our study adopts the existing method that feeds the conditional time-invariant data after running the sequential layers. This approach is discussed by Wang *et al.* in their literature. Our alumni dataset includes both the dynamic and static data. The dynamic data contains the past ten year's time-series data. The current static data may differ from the data 10 years ago, including age, education, marital status, etc., which may affect the giving decision in the coming years. Therefore, adding the most current demographic data at the end of the ten-year sequential layer is the most reasonable choice for the static and dynamic fusion model design. We compare the performance of models from three groups, the traditional model with the static alumni data, the sequential with dynamic data, and the sequential model combining

the time-variant and time-invariant features. The model architecture is illustrated in Figure 3.4.

Finally, the impact of database size on the effectiveness of deep learning model should be addressed. Many machine learning models, especially deep learning, a subset of machine learning, perform better as the dataset size gets bigger. Abdulraheem *et al.* [35] demonstrated in their study how the database size affects the supervised neural network models. Deep neural networks have many parameters to learn, which indicates many iterations to find the optimum values. Running a large number of iterations on a small dataset, therefore, can result in overfitting. By using a large dataset, we can avoid overfitting and generalize better on new dataset.

3.2 Model Architecture

In this section, we discuss the theories of the algorithms we utilized for our sequential models as well as brief description of the traditional model architectures.

Traditional vector-based model takes feature vectors $f_1, f_2, f_3 \ldots . f_n$ where f_n stands for feature n. The traditional approach requires feature engineering to identify the time-dependent donation history in the model. In contrast sequential models enable us to capture donor's historical donation behaviors in time sequence. [38]. The time-series features in the static vector-based model are independent at each data point relative to others, which results in loss of the temporal coherence. On the contrary, sequential models like Recurrent Neural Network (RNN) and a family of the sequential models consider the time-sequence of inputs. The family of sequential models such as RNN, LSTM and GRU are often used for machine translation [13], speech recognition [36], stock forecasts [46], and autonomous driving [41]. Our study goal is to capture the donor's time-series financial contribution by comparing the performance against the traditional vector-based machine learning models. In the following section, we introduce the main concept of the algorithms used in our study.

3.2.1 Recurrent Neural Network

To understand Long Short Term memory (LSTM) and Gated Recurrent Unit (GRU) better, Recurrent Neural Network (RNN) is reviewed first. RNN [49] is a family of artificial neural networks whose structure is similar to a feedforward neural network. In the feedforward neural network, inputs are fed from layer to layer in one

direction, and the previous information is not considered in the model. In contrast, the information in RNN flows recursively from layer to layer. It allows the state of the model to be influenced by its previous state. RNN takes a sequence of each feature by the length of time (T).

$$h_t = f(w_x x_t + w_h h_{t-1} + b) \tag{3.1}$$

where

$$h_t = hidden\ state\ at\ time\ t \tag{3.2}$$

$$X = (x_1, x_2, x_3, \ldots\ldots x_T) \tag{3.3}$$

$$x_t = input\ at\ time\ t \tag{3.4}$$

The first equation is calculated based on the previous hidden state h_{t-1} and the input x_t at the time t. w_x is the parameter weight associated with input and w_h is the one associated with the hidden state. The weights are shared across all time steps. The function f is a nonlinear activation function such as tanh, ReLU, and sigmoid [43]. The first hidden state is normally initialized to random values close to zero. o_t is the output at time t. Softmax is used to obtain a vector of probability for binary or categorical values, .

$$o_t = softmax(w_o h_t) = \frac{1}{\sum_{j=i}^{K} exp(w_o^T h_t)} \tag{3.5}$$

Softmax is a function that converts a vector of real values to the same dimensional vector (K-dimensional) that ranges from 0 to 1 and adds up to 1 [11].The values received from the output layer are normalized by the sum of the exponentiated values obtained at all the output nodes and turn them into a probability. An output probability for each node is categorized into a particular class by selecting a class with maximum probability. There are various softmax alternatives, Taylor softmax, soft-margin softmax, and SM-Taylor softmax, but we used the regular softmax function for our study. While a traditional neural network has different parameters at each layer, an RNN shares the same parameters, w_x, w_h, w_o for input, hidden layer, and output across all steps.

One issue with RNN is the vanishing/exploding gradient descent problem, limiting looking back in time [37]. The problem arises when it back-propagates through time. In the propagation process, the cost function goes through the layer backward all the

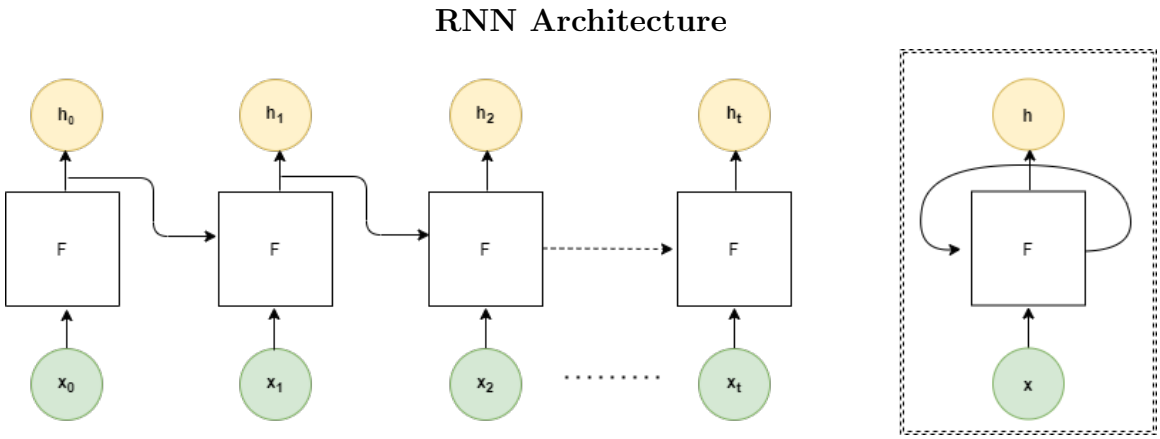

Figure 3.1: Unrolled and Rolled RNN Architecture where F is the activation function - insprired by Christopher Olah [33]

way through the time to update the weights. When the weights W are initialized to close to zero and multiply x_t, $x_{(t-1)}$, $x_{(t-2)}$,... by the close-to-zero weights, the gradients become less and less at every time step going backward. It means it is harder to update the weights and takes longer to get the final result. On the other hand, if the gradients are big, they get bigger and bigger as propagating in time.

In the exploding gradient case, the solution is to penalize or artificially reduce the gradient or put a maximum limit on a gradient. The popular solution for the vanishing gradients problem is the Long Term Short Term Memory Networks (LSTM) discussed in the next section.

3.2.2 Long Short Term Memory

The Long Short Term Memory (LSTM) Networks is a special type of RNN, which is capable of learning long-term dependencies (over 1000 time steps) [54]. LSTM stores information outside of the recurrent network in a structure called a gated cell. It works like a computer memory where information is stored, read, and written. The gated cell decides what to store, when to read, and write by opening and closing the gate [33];the standard RNN does not have this gated cell. In standard RNN, each sequential layer has a simple structure with a single activation function.
LSTM has three gates at each time step, forget gate, input gate, and output gate that decide what and how much information goes through.

- Forget gate layer (f_t) (eq. 3.6) decides which parts of the old output, $h_{(t-1)}$

should be kept or thrown away. It is a sigmoid function (σ) calculated with $h_{(t-1)}$ and X_t values where $h_{(t-1)}$ is the old output and X_t is the new input. The calculated values by sigmoid function fall between 0 and 1. 1 represents to keep the information, and 0 to get rid of the information in the values in the cell C_{t-1}. The values of the forget gate layer f_t is multiplied by the previous cell state C_{t-1} in eq. 3.9 to determine which previous state is to be stored in the state C_t.

- The next gate layer has two layers, which decide how much new information should be stored in the cell state C_t. The first layer, the input gate layer i_t calculated by the sigmoid function determines which values should be updated (eq. 3.7) The tanh layer creates all the candidate values to update (eq. 3.8). These two values derived from the sigmoid and tanh functions are multiplied $i_t * \overline{C}_t$ to be added to the updated values by multiplying the old cell state $C_{(t-1)}$ and the forget gate value f_t (eq. 3.9).

- The last gate layer is an output layer. First, the sigmoid function derives which part of the cell state C_t should be outputted (eq. 3.10). Then to generate the output at the time t, the cell state C_t is gone through the tanh function and multiplied by o_t derived by the sigmoid function (eq. 3.11)[33][55].

$$f_t = \sigma(W_f x X_t + W_{fh} h_{t-1} + b_f) \tag{3.6}$$

$$i_t = \sigma(W_{ix} X_t + W_{ih} h_{t-1} + b_i) \tag{3.7}$$

$$\overline{C}_t = tanh(W_{\overline{C}x} X_t + W_{\overline{C}h} h_{t-1} + b_{\overline{C}}) \tag{3.8}$$

$$C_t = f_t C_{t-1} + i_t \overline{C}_t \tag{3.9}$$

$$o_t = \sigma(W_{ox} X_t + W_{oh} h_{t-1} + b_o) \tag{3.10}$$

$$h_t = o_t * tanh(C_t) \tag{3.11}$$

3.2.3 Gated Recurrent Unit

Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU) were developed to deal with the vanishing and exploding gradient issues. GRU was introduced by Cho *et al.* [27] in 2014. LSTM and GRU have more similarities than differences. The key differences between LSTM and GRU are:

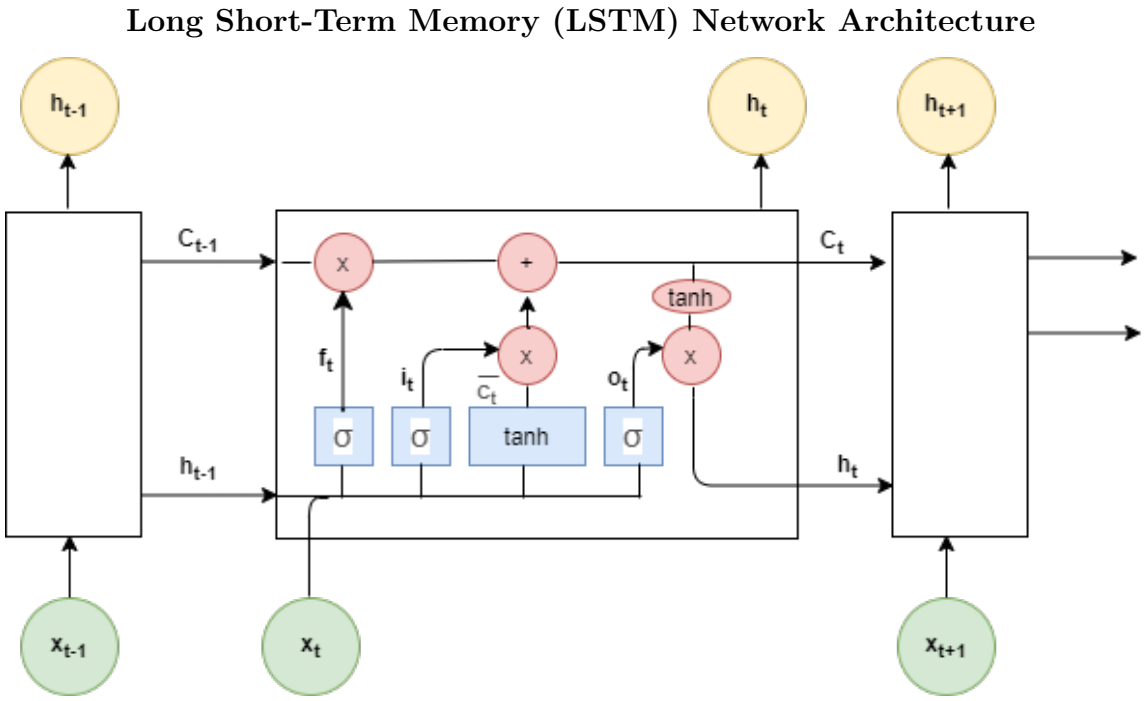

Figure 3.2: Long Short Term Memory (LSTM) Networks is a special type of RNN. Unlike RNN, LSTM is capable of learning long-term dependencies (over 1000 time steps) [25]. Figure was inspired by Christopher Olah [33]

1. An LSTM has three gates (forget gate, input gate, and output gate) while a GRU has only two gates (reset gate, and update gate).

2. GRU uses fewer training parameters, and therefore, GRU is computationally more efficient.

3. Because GRU does not have an additional cell state, the memory operates directly on the hidden state.

The GRU cell combined the forget gate and input gate into an update gate. Cho *et al.* [27] evaluated vanilla RNN, LSTM, and GRU's performance and demonstrated that both LSTM and GRU perform better than vanilla RNN.

$$r_t = \sigma(W_{rx}X_t + W_{rh}h_{t-1} + b_r) \tag{3.12}$$

$$z_t = \sigma(W_{zx}X_t + W_{zh}h_{t-1} + b_z) \tag{3.13}$$

$$\overline{h}_t = tanh(W_{h\overline{x}}X_t + W_{\overline{h}h}(r_t h_{t-1}) + b_{\overline{h}}) \tag{3.14}$$

$$h_t = z_t\overline{h}_t + (1 - z_t)h_{t-1} \tag{3.15}$$

The GRU has two gates: the reset gate and the update gate. The reset gate controls how much information from the previous hidden layer to forget. It is activated by a sigmoid function whose values are between 0 and 1, as in LSTM, which indicates the importance of information (eq. 3.12). The same mathematical equation is used for the update layer z_t to derive what new information to be passed on and what past information to discard (eq. 3.13). The weights W_{zx} in the update layer are different from the weights W_{rx} in the reset layer. The previous hidden state h_{t-1} is controlled by r_t. When r_t is closer to 1, the formula becomes the same as the original RNN, $h_t = tanh(W_{hx}X_t + W_{hh}h_{t-1} + b_h)$ (eq. 3.1). When r_t is closer to 0, it becomes like Multiple Layer Perceptron (MLP), by turning any previous hidden state to defaults. By taking a tanh it ensures that the values remain in the interval of (-1, 1). This state is called a candidate hidden state $\overline{h}_t$. Lastly, the previous state and candidate hidden state are multiplied by the update gate state. When the update gate z_t is closer to 1, the new latent state approaches to the candidate hidden state. In contrast, when z_t is closer to 0, the previous hidden state is retained while the information from any information from X_t is ignored [55] [31].

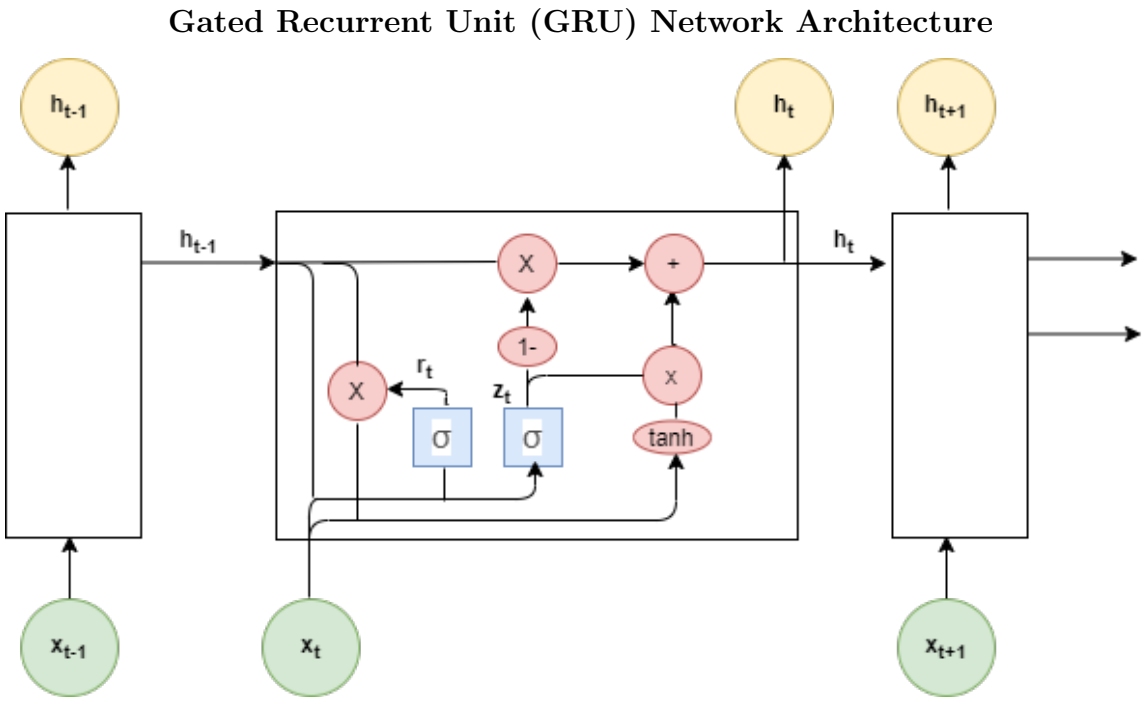

Figure 3.3: GRU does not use a memory unit $_t$ and directly operates on the hidden state. GRR is computationally more efficient than LSTM because GRU uses fewer training parameters. It exposes the full hidden content without any control.

3.2.4 Neural Network

A neural network is the model that mimics the structure and function of the human brain. It has three layers of nodes called input, hidden, and output. The input layer contains input variables, while the output layer has target variables. A Neural network can have more than one hidden layer, but adding a hidden layer increases code complexity and processing time. All the input nodes are connected to every hidden node, which is also connected to every output node. By training a neural network, the weight values are optimized to minimize the error between true values and estimated values.

3.2.5 Support Vector Machine

SVM can be used for both binary and multi-class classification. A binary SVM classifier works by identifying the optimal hyperplane that divides the data points into two classes. SVM selects the best hyperplane with a maximum margin, the distance between the classifier and support vector.

3.3 Data preprocessing and model architecture

Many machine learning models use either static(time-invariant) or dynamic(time-variant) data type. In real life scenario, however, there are many cases where both static and dynamic features are present in the data. In our bimodal model, we combined both types of data. One of our contributions is to examine if combining static and dynamic features will improve classification performance. In this section, we demonstrate how we combine both static and dynamic features in the sequential model.

3.3.1 Bimodal model

Our alumni data contains both static and dynamic data. The static information is the alumni's demographic information such as age, marital status, education, etc., while dynamic data includes time series donation, communication information (email open/clickthrough), and event attendance information. In this study, unimodal means using either static or dynamic data separately, while bimodal means using both static and dynamic data. We utilize the unimodal neural network with backpropagation and

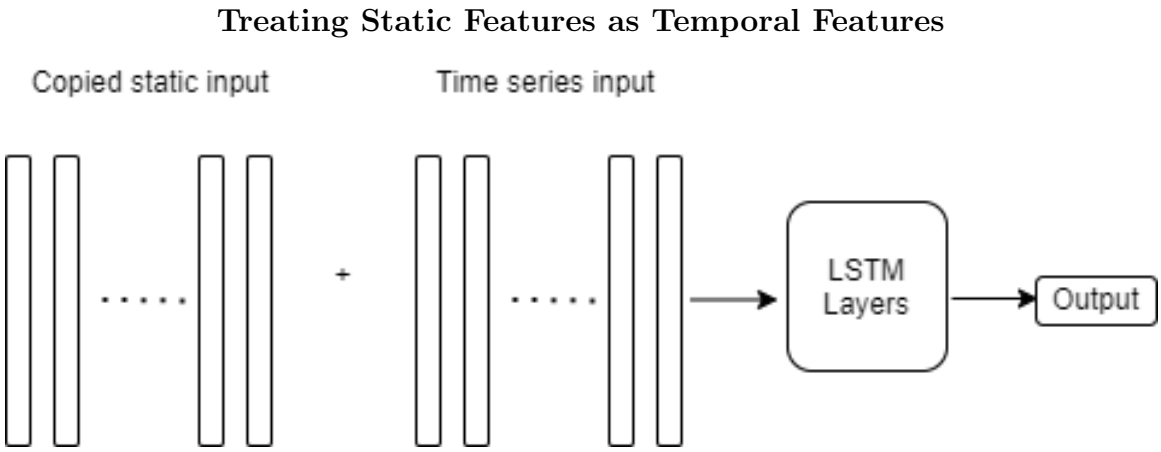

Figure 3.4: Method 1 - Copying t times of static features and adding them to the time series features

SVM as baseline models, which treat all data as static, i.e. both static demographic data and dynamic giving information. Traditional models are discussed in detail in Subsection 3.3.3. In contrast to the traditional model, our experiment model treats the static demographic data as conditions to the sequential LSTM model instead of mixing time-series data and non-time series data in the sequential layers. The time-invariant demographic data is introduced outside of LSTM layers and combined with LSTM output. We examine if adding the conditional demographic data affects model performance. Four possible methods of merging static features and dynamic features summarized below.

1. Treat the static features as dynamic features and include them in the dynamic time-variant features at each stage. (Fig 3.4)

2. Concatenate the static features with the output of the hidden state at the last time step. (Fig 3.5)

3. Apply the static time-invariant feature to initialize the first hidden layer (t = 1) and concatenated with the dynamic features at the following stages (t > 1) (Fig 3.6)

4. Concatenate the static features with the hidden state at each time step. (Fig 3.7)

We adopted the second method (Fig 3.5) for our model comparison because it is reasonable to assume that the latest demographic information, such as age, education,

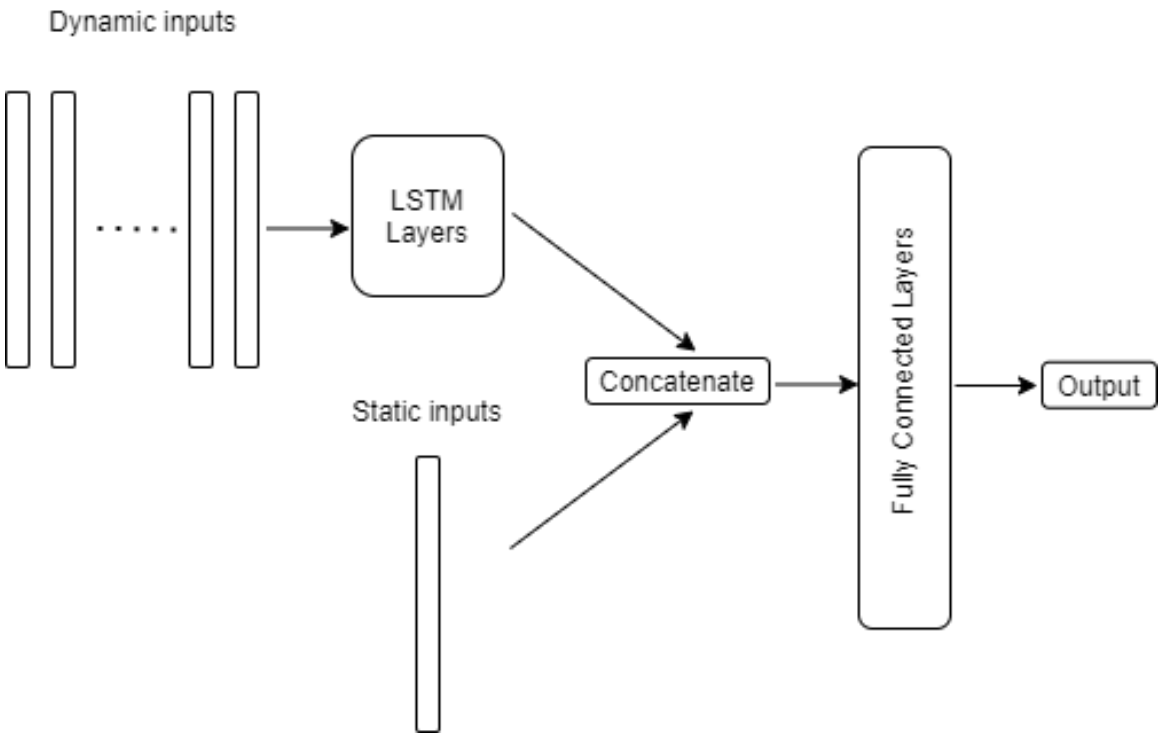

Figure 3.5: Method 2 - Concatenating the static features with the output of the hidden state at the last time step.

and family information, affects the alumni giving decision for the coming years instead of utilizing demographic information that is ten years old. Therefore, we add the static data at the end of the LSTM time step. The mathematical expression of the LSTM hidden layer calculation for the third method is expressed as follows.

$$h_t = 1(t < T) * \sigma_t * tanh(C_t) + 1(t = T) * tanh(W_{con} * c_i + b_{con}) \tag{3.16}$$

where c_i = Conditions (static demographic features)
The hidden state is set to the condition, c_i, the static data at the last stage, t=T. The conditional hidden state, $\tanh(W_{con} * c_i + b_{con})$ is added to the end of LSTM time-series layer, t=T. $\sigma_t * \tanh(C_t)$ is the computation of hidden states, h_t in LSTM, equivalent to the equation (3.11) while the other mathematical equations (3.6) to (3.10) remain the same.[8]

In the study, we examined the traditional unimodal model and the bimodal model with time-variant and time-invariant features combined.

Initialize static input and concatenate with the dynamic input to run LSTM

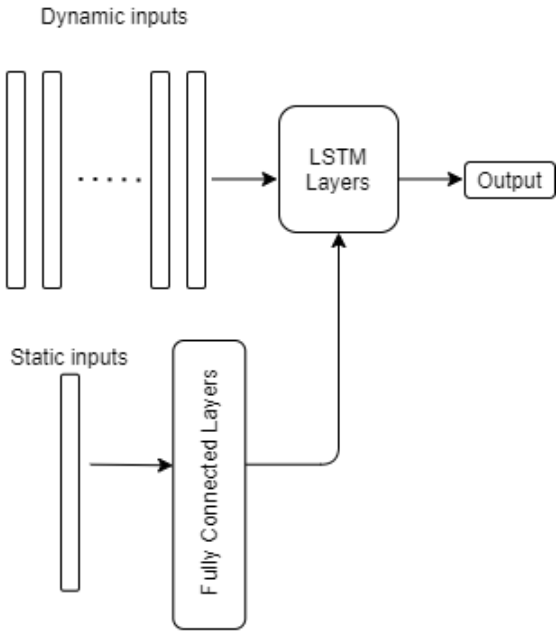

Figure 3.6: Method 3 - Apply the static time-invariant feature to initialize the first hidden layer (t =1) and concatenated with the dynamic features at the following stages

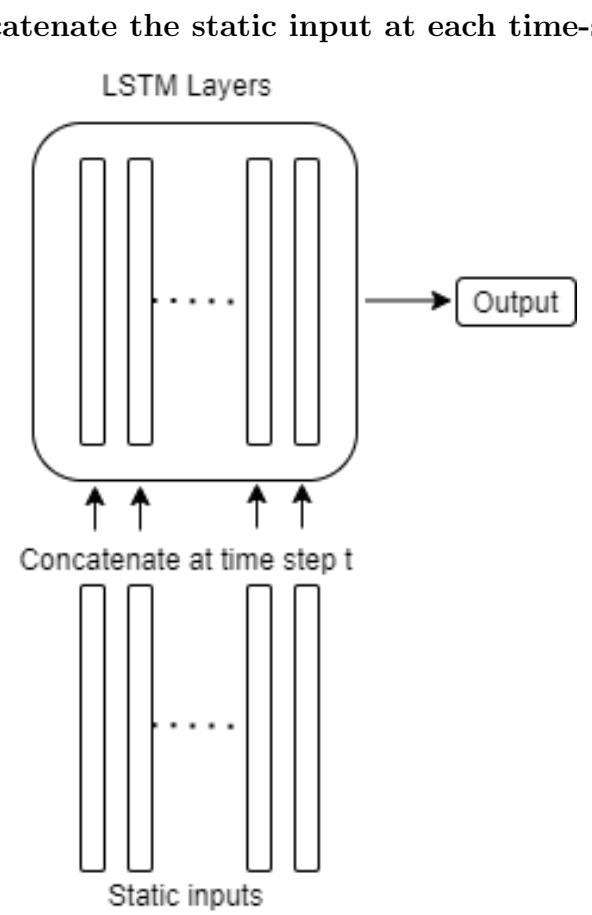

Figure 3.7: Method 4 - Concatenate the static features with the hidden state at each time step

3.3.2 Data Preparation

Muralidharan (2010) [31] researched the importance of data cleansing and processing. Data cleansing and processing involve removing errors, synthesizing missing values, creating categorical variables, adjusting outliers, and standardizing/normalizing variables. This section illustrates the methods and techniques we use for data preparation for the model.

Missing Data

Missing values occurs for a various of reasons. The large percentage of non-alumni records lacks demographic and education information. By contrast, this information is collected when alumni are enrolled in the University; therefore, only 2% out of all alumni records lack demographic information. For this reason, the non-alumni data was removed from our study. The constituent's wealth information was estimated by the Canada census data released every four years. If the records do not have a full address, including the correct postal code, the estimated wealth values will not be generated. Missing data can result in loss of efficiency, and higher bias. We used descriptive analysis to understand the data and data issues, including missing data, data imbalance and skewed data

To measure the importance of features that have a larger occurrence of missing values, we conducted a correlation matrix analysis on all features and target variables. The correlation matrix table (Table 3.1) does not include features of less than 10% of correlation. After close examination, the best features are selected from the correlation matrix and RFE feature selection described later in Table 3.4. This process analyzes how strong features and target variables are associated, and helps select better-correlated variables to the target variable for better model performance. If the features have high frequency of missing values, yet have a higher correlation with the target variable, we include those features by synthesizing missing values. The target variable in our study is the largest gift amount out of all the past donations for each constituent. The analysis revealed that most of the feature variables have a thin correlation score. Among them, some features indicate a relatively higher correlation, including age, wealth, years donated, loyalty score, click-through count, number of relationships with organization and individuals, board member, event attendance, and gender.

Feature Variables	Description	Corr Score
BIO-Age	Alumni Age	0.136199
GIO-Gender Category	Alumni Categorical Gender	0.130422
EDU-Law Degree Count	No. of Law Degrees Received	0.151982
EVT-Total Registered	No. of Even Registration	0.343171
EVT-Vikes Athletics	No. of Athletics Event Registration	0.190436
EVT-Homecoming	No. of Homecoming Event Registration	0.219283
REL-Board Member	No. of Board Volunteer	0.152661
REL-Child	No. of Children in the Record	0.137688
REL-Sibling	No. of Sibling In the Record	0.152431
REL-Relative	No. of Relatives in the Record	0.132748
REL-Spouse	Spouse Information in the Record	0.161355
REL-Org Relations	No. of Organizations Linked to the Record	0.2662
REL-Ind Relations	No. of Individuals Linked to the Record	0.294414
REL-Alumni in Family	No. Alumni in the Family	0.185035
BBNC-Email Received	No. of Emails Received	0.166905
BBNC-Opened	No. of Emails Opened	0.201846
BBNC-Clicked Through	No. of Click Through	0.182464
GIFT-Loyalty Score (Weighted)	Weighted Donor Loyalty Score	0.376365
pca 1 – Wealth data	PCA of Wealth Information	0.142728
Loyalty Years	No. of Years Donated	0.576628

Table 3.1: Correlation between Feature Variables And Target Variable (Log of Largest Gift Amount)
NOTE: Features with less than 10% correlation coefficient are excluded. NOTE: Loyalty Score is a proportion of total years donated divided by the years from the first and last year donated. The loyalty Score (Weighted) is adjusted by the frequency of donations.

The correlation of alumni's age and largest gift amount (target variable) is not strong; however, the relationship between alumni's age and average largest gift amount clearly indicates a positive relationship as Figure 3.8 illustrates. In Figure 3.9 and 3.10, the warm colors indicate a higher donation amount. As in the charts, the number of warm colour circles (pink-red) increases as the age range of the largest gift gets higher, meaning that the average gift size increases as donors get older. Although the largest gift and age do not show a strong correlation , these charts reveal that the donor's age is one of the important factors that influence philanthropic decisions. Another feature that shows the overall positive relation is wealth PCA, as shown in Figure 3.11.

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of possibly correlated variables into a set of values of linearly uncorrelated variables, called principal components.[14] The first principal component has the largest variance of variables, and each succeeding component has the highest variance under the restriction that the component is orthogonal to the preceding component. Our wealth data is estimated by a census-tract level median household discretionary income, median household disposable income, median house value, median liquid asset, etc. Those values are highly correlated each other, and the redundant information needs to be removed from the dataset. Although PCA values are not highly correlated with the target variable, figure 3.13 shows a positive correlation between gifts and PCA values, indicating that wealth affects the amount of financial contribution. The prospect research at any non-profit organization performs wealth screenings to determine a donor's financial giving capability. We included wealth data in our model because of the reason discussed above and the fundraising business practice. The first principal component (pc) variance is 7.006, 1.498 for the second, and 0.6901 for the third on the normalized wealth data. The difference between the first and the second principal components is considered significant for the normalized wealth data. Therefore, we included the first principal component for the model based on the correlation analysis and standard fundraising practices.

The missing value issue is common to machine learning. As Table 3.3, 3.4 and 3.5 indicate, a high percentage of missing values is observed in features like age, employment, and contactable address compared to other features. The postal codes are used to match the average wealth data with the dissemination area of constituent's address. About 13% of all records, 2% of alumni records, and 70% of non-alumni records have missing age values. About 12% of alumni records and 10% of non-alumni records don't have postal codes. As discussed in 3.2, a large proportion of education and demographic data is non-existent in the non-alumni records; in contrast, the alumni personal information is collected when they enrolled in the program. The alumni personal data is transferred to the UVic alumni/fundraising database once they graduate from UVic. Because non-alumni data has more missing information including age, postal code, education information, we separate the alumni and non-alumni dataset and apply the models only to the alumni data for our study. 2% of missing age values among alumni records are first estimated by the graduation year. The alumni demographic data of the Normal School and Victoria College was either not collected or

MICE Imputation Steps	
Step 1 :	Impute all missing values except the first column by using column mean values.
Step 2 :	Run the Linear Regression on the first feature column as target variables against other feature columns as independent feature variables.
Step 3:	Update the missing values with the predicted values. Run the Linear Regression on the next column as target variables against other columns including the first column updated with the predicted values. Continue this process until all the missing values are imputed.

Table 3.2: MultiMultivariate Imputation By Chained Equations (MICE) Imputation Steps

has been lost over the years. The age and wealth data are imputed using the missing value estimation method.

Multiple Imputation of Chained Equations (MICE)

Among the methods to mitigate the missing data issues, we applied Multiple Imputation of Chained Equations (MICE) [21] to impute the missing age and wealth data. MICE uses a chain of regression equations that imputes variables with missing data one by one. The first variable with missing values is imputed on all other variables. Missing values are then replaced by the imputed values. Then the next variable is regressed on all other variables, including the variable whose missing values are replaced. This process continues until all missing values are estimated. (Table 3.2) The accuracy of the imputations depends on how independent variables are correlated. If independent variables are completely independent, MICE does not yield accurate imputed values.

Class Imbalance

A common problem with machine learning classifiers is a class imbalance, where some classes have a markedly higher number of observations than other classes [30]. The class imbalance problems can be a critically adverse effect on the classification models. Examples include medical diagnosis, fraud detection, and network intrusion detection [40]. It affects convergence during training and generalization on the test

set [40]. The dataset used for this study consists of a disproportionately high number of non-donors. There has been lots of research for class imbalance, and we explore one of the techniques for the class imbalance problem in our study.

The class imbalance solutions include re-balancing training dataset [1], cost-sensitive method [53], and one-class classification [29]. The re-balancing training dataset technique has two methods: over-sampling and under-sampling. Due to the relatively small size of alumni data, we employed over-sampling to synthesize the minority classes (class 1, 2, 3) by using SKlearn SMOTE package. Over-sampling is to create new synthetic data points that are similar to the existing data point. SMOTE generates a synthetic minority observation considering k-nearest-neighbors close to the existing minority observation. The SMOTE over-sampling procedure picks one minority class data point, selects one of the K-nearest minority class neighbors, and creates synthesized data points based on the pattern, $Z = X_0 + w(X - X_0)$ where w is a uniform random variable. The procedure repeats the same process to synthesise minority points.

Normalization and Transformation

The purpose of normalization and transformation is to make the data more reliable for model computation in the presence of skewed distribution, nonlinearity, and multicollinearity. Table 3.12 suggests that 76.5% of alumni are non-donors, while a mere 0.02% of alumni donors have given 67.7% of the total donation amount from all alumni. The donors whose Largest Gift amount is greater than $10,000 is only 7.69% of all alumni donors. The median of the largest gift amount is $50.00, and the maximum amount is $5,000,000 among alumni donors. The donors are more on the left side of giving distribution, which means a higher concentration of small donors. The gift amount is, in contrast, negatively skewed, as shown in Figure 3.13, which suggests that big donors are outliers (Figure 3.14). Because we are looking for donors, including big donors, i.e. the outliers are critical for the analysis as well. Among various techniques to mitigate the skewed distribution, we applied a logarithm on the gift related features, log(x) to transform the skewed distribution into the normal distribution as shown in Figure 3.11. The gift related features include lifetime gift amount, average gift amount, largest gift amount etc., whose range is large. The rest of the features are normalized between 0 and 1.

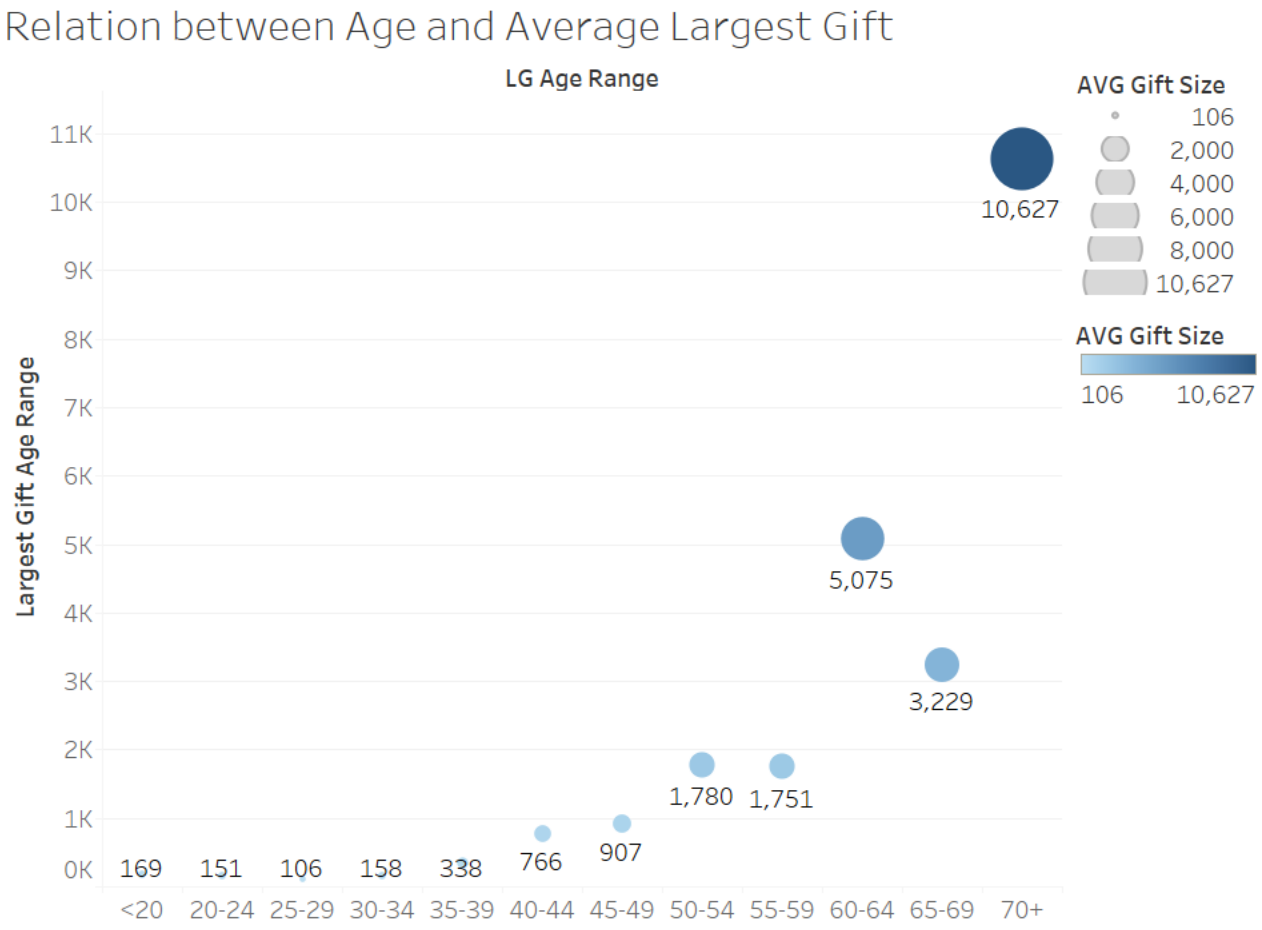

Figure 3.8: The Age range when the largest gift was made is positively related the average donation amount for each age range.

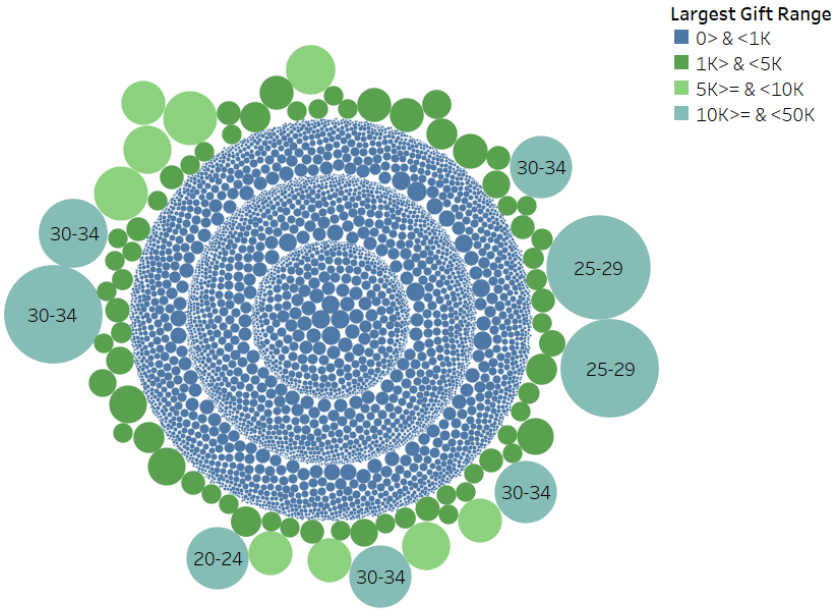

Figure 3.9: Donor Distribution Among Age range 20-34
The size and the colour of the circle denote the size of the gift amount. The bigger the circle and the warmer the colour, the larger the gift amount.

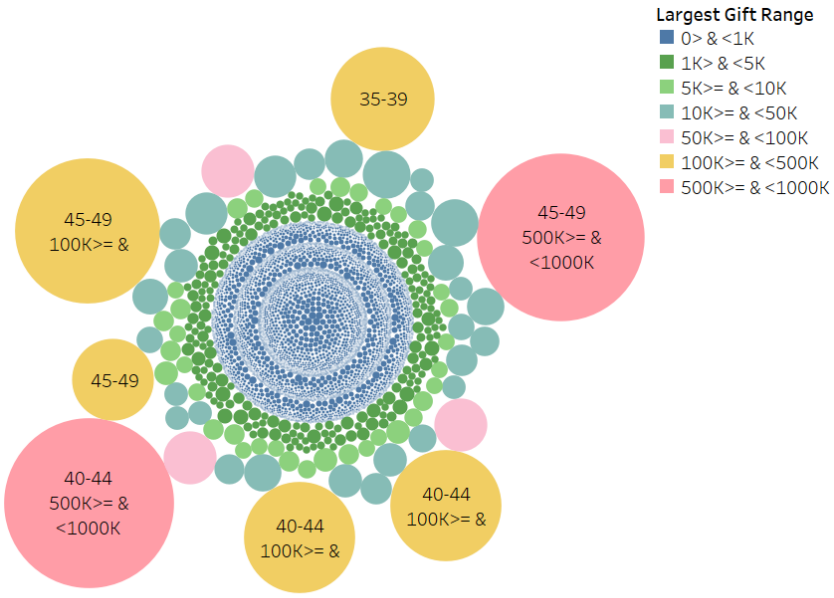

Figure 3.10: Donor Distribution Among Age range 35-49
The size and the colour of the circle denote the size of the gift amount. The bigger the circle and the warmer the colour, the larger the gift amount.

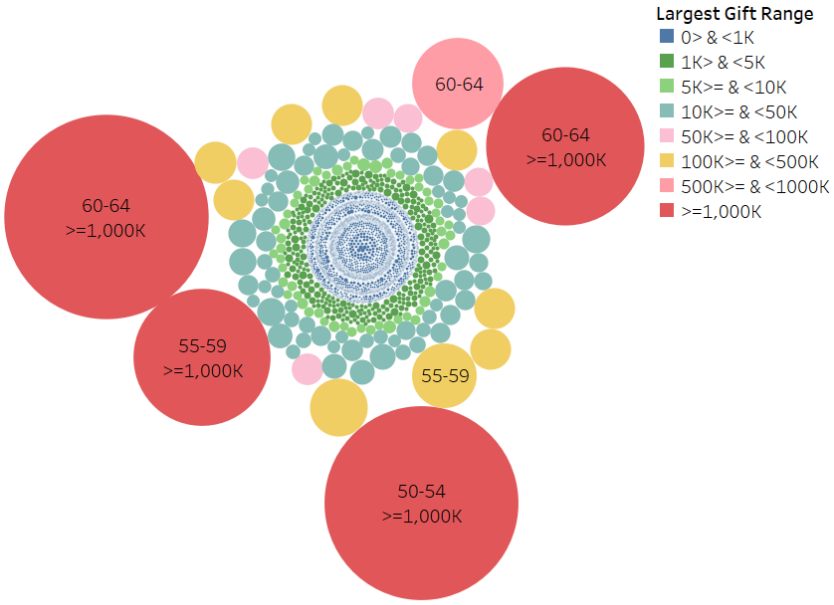

Figure 3.11: Donor Distribution Among Age range 50-59
The higher the age range is, the more visible presence of warm coloured circles of the large-gift donors.

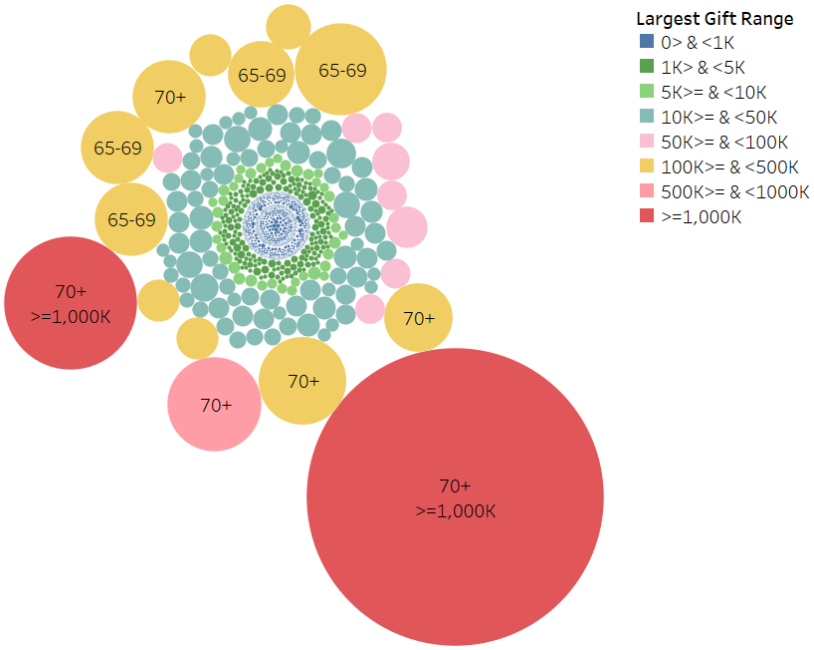

Figure 3.12: Donor Distribution Among Age range 50-59 & 60+
The higher the age range is, the more visible presence of warm coloured circles of the large-gift donors.

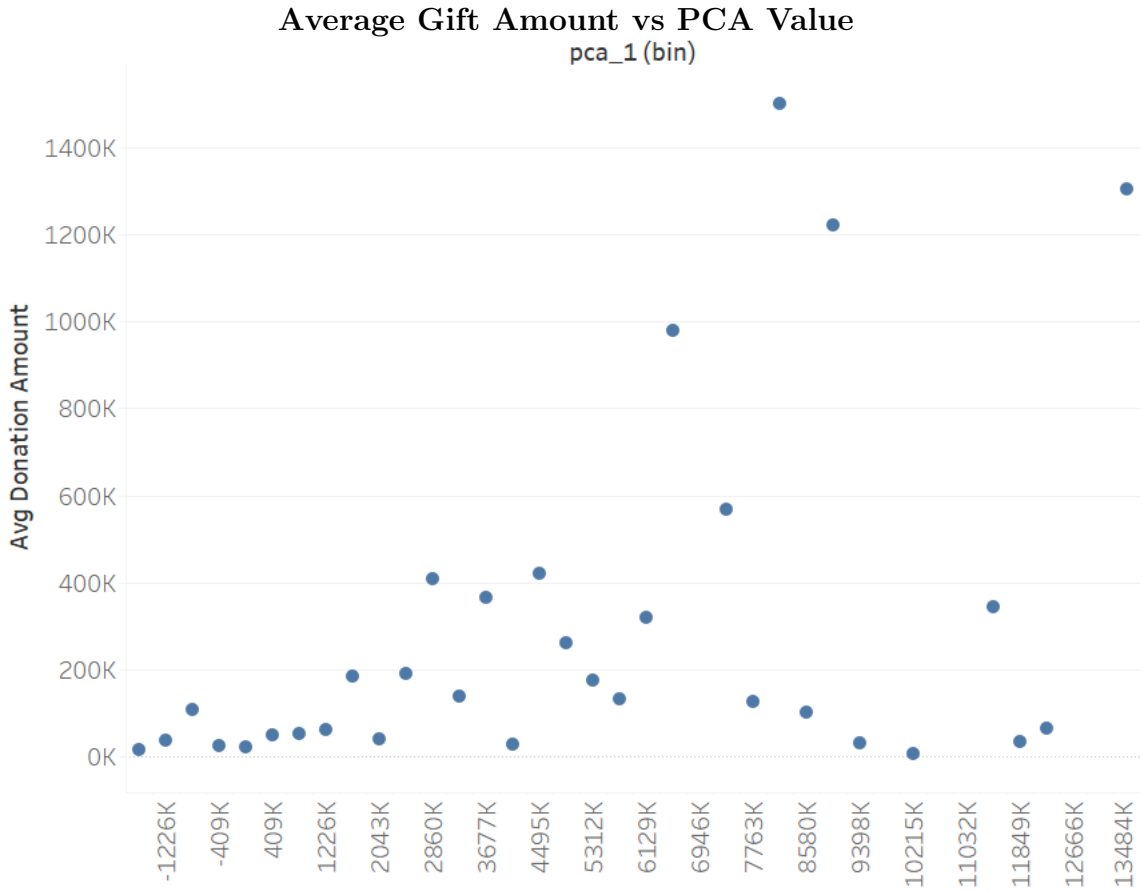

Figure 3.13: Average gift amount increases as the PCA value increases.

Largest Gift Distribution - Count

Largest Gift Range	Count	% of LG
Non-Donor	93,190	76.50%
0> & <1,000	26,425	21.69%
1,000> & <5,000	1,446	1.19%
5,000>= & <10,000	275	0.23%
10,000>= & <50,000	358	0.29%
50,000>= & <100,000	44	0.04%
100,000>= & <500,000	45	0.04%
500,000>= & <1,000,000	7	0.01%
>=1,000,000	25	0.02%

Figure 3.14: Distribution of Donors among Live Contactable Alumni

LG Gift Distribution - Sum

Largest Gift Ran..	Sum	% of Total S..
Non-Donor	0	0.00%
0> & <1,000	3,104,126	3.47%
1,000> & <5,000	2,450,413	2.74%
5,000>= & <10,0..	1,684,645	1.88%
10,000>= & <50,..	7,373,357	8.24%
50,000>= & <100..	2,601,567	2.91%
100,000>= & <50..	7,750,844	8.66%
500,000>= & <1,..	4,018,784	4.49%
>=1,000,000	60,468,681	67.60%

Figure 3.15: Distribution of Gift Amount Among Alive Contactable Alumni

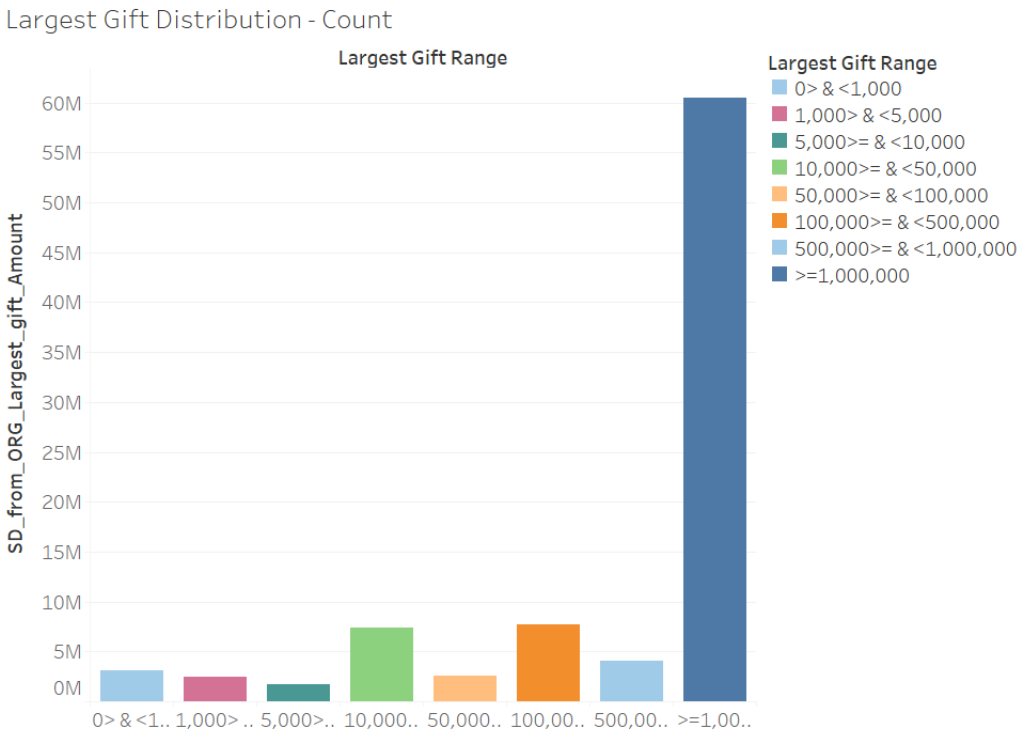

Figure 3.16: Negatively Skewed Largest Gift
The number of big donors are small, but the total amount of large donation is concentrated in the largest range of the gift amount.

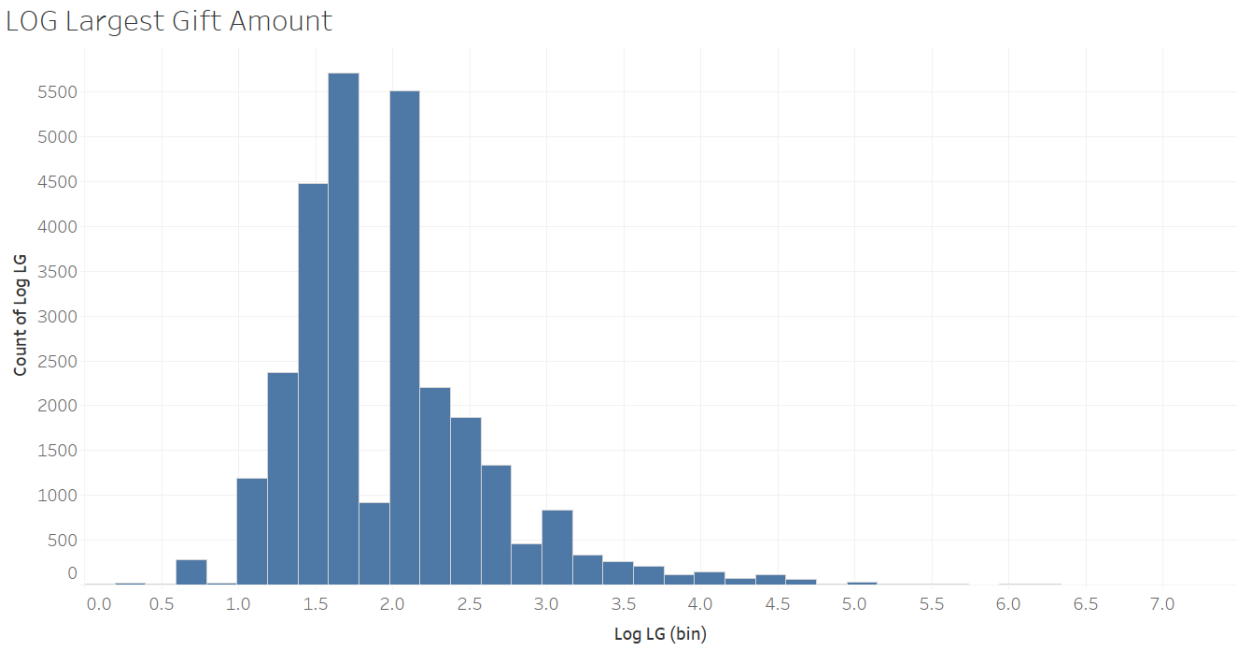

Figure 3.17: Log Transformation
Positively skewed distribution is transformed to normally distributed distribution after taking logarithm of the largest gift values.

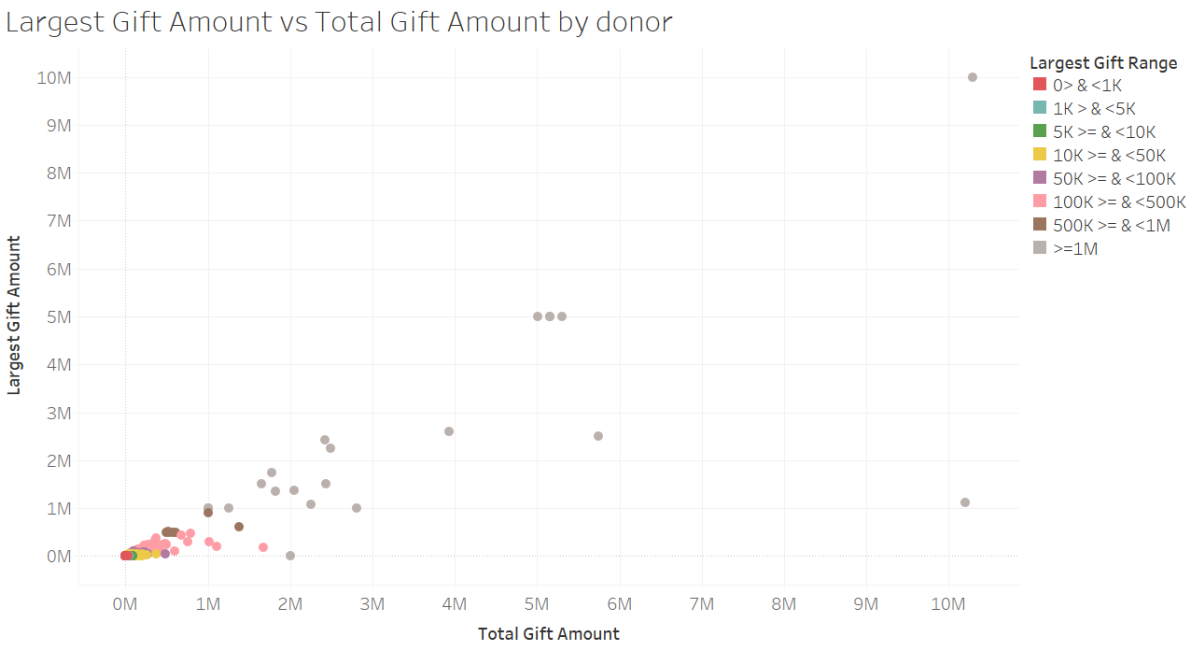

Figure 3.18: Largest Gift Amount vs Total Gift Amount by Alumni
Most of donors are small donors that are concentrated under $1M and big donors are small in numbers and considered to be outliers.

Stat	Amount
Min	$1.00
Max	$5,000,000.00
Median	$50.00
Skewness	57.62955

Table 3.3: Overall Donor Largest Gift Statistics
The overall gift amounts are positively skewed with the median gift value being $50.00.

3.3.3 Model

This section illustrates the training methods of the baseline models and the proposed models.

For our study, we used Python for the programming language. In addition, we used Python libraries including Numpy, Pandas, Keras and Scikit-learn for processing, model training/testing and assessment.

Traditional Model

For the traditional model, we selected Neural Network and Support Vector Machine (SVM). The baseline Neural Network Model is structured with two Relu and sigmoid layers and softmax for the multi-class prediction. The hyper-parameters for the SVM model are optimized by applying Grid-search with ten-fold cross-validation.
The objective of the sequential model is to classify future giving amounts by using alumni's historical data including donation amount, email open/click through. Because of the time-invariant assumption for the traditional machine learning model, the traditional model and sequential models can not employ the same model architecture. All the time-variant features including donation, event attendance, and email communication numbers are taken from the fixed time-window and used as time-invariant features in the traditional models.
As discussed in 3.2, the traditional model requires feature engineering to identify the time-dependent donation history in the model. Feature engineering is used to optimize the model accuracy for the traditional models. For the performance comparison, in addition to the features selected by the feature selection method, the donation features are included in the traditional model:

- Years from the first $1,000 plus gift being made
- Years from the first gift being made

- Years from the last gift being made
- Years from the first gift until the first $1000 plus gift was made

These factors are considered for the standard prospect research practices to assess how prospective donors are ready to give. If the last gift was made many years ago, it is less likely to receive donations from this kind of lapsed donor. By including these constructed variables based on the existing variables.

We used Logistic Regression - Recursive Feature Elimination [31], correlation matrix, and subject expert opinions to determine which features should be included in the experiments. The Recursive Feature Elimination (RFE) selects features backwardly. This technique is used in conjunction with a model; however, not all models can be paired with RFE. The models that can be paired with RFE are multiple linear regression, logistic regression, and linear discriminant analysis. We employed logistic regression, together with RFE, to computer the importance score for each feature. After removing the features with a lower score, the model is rebuilt, and the importance scores are computed again. This process is repeated until it optimizes the performance criteria. Among those selected features, we divided them into the time-variant and time-invariant features. The results of feature selection using RFE are depicted in Table 3.4. The features from the Correlation Matrix results (Table 3.1), the logistic regression with RFE, and the features based on the domain expert's suggestions are integrated into the traditional model. Those techniques are also applied on the alumni's demographic features of the sequential models.

The dataset is split at the ratio of 3:1; one dataset is used for training and the other dataset for testing by applying 10 fold cross-validation. The largest gift is set as a multiclass target with four classes based on its gift range. The feature variables were taken from the data up to 2019, and the target variable is from 2020 to predict the future gift range. The target classes have four groups, $0, greater than $1 and less than $1,000, greater than $1,000 to less than $10,000, and greater than $10,000 for the one-year time shifted donation data.

Sequential Model - LSTM & GRU

Our study focuses on introducing time dependency into the fundraising model and comparing it with the traditional time-independent model. We selected variations of the sequential architectures, namely stacked LSTM/GRU, Bidirectional model, Time-distributed model and LSTM/GRU with condition. We employed LSTM and

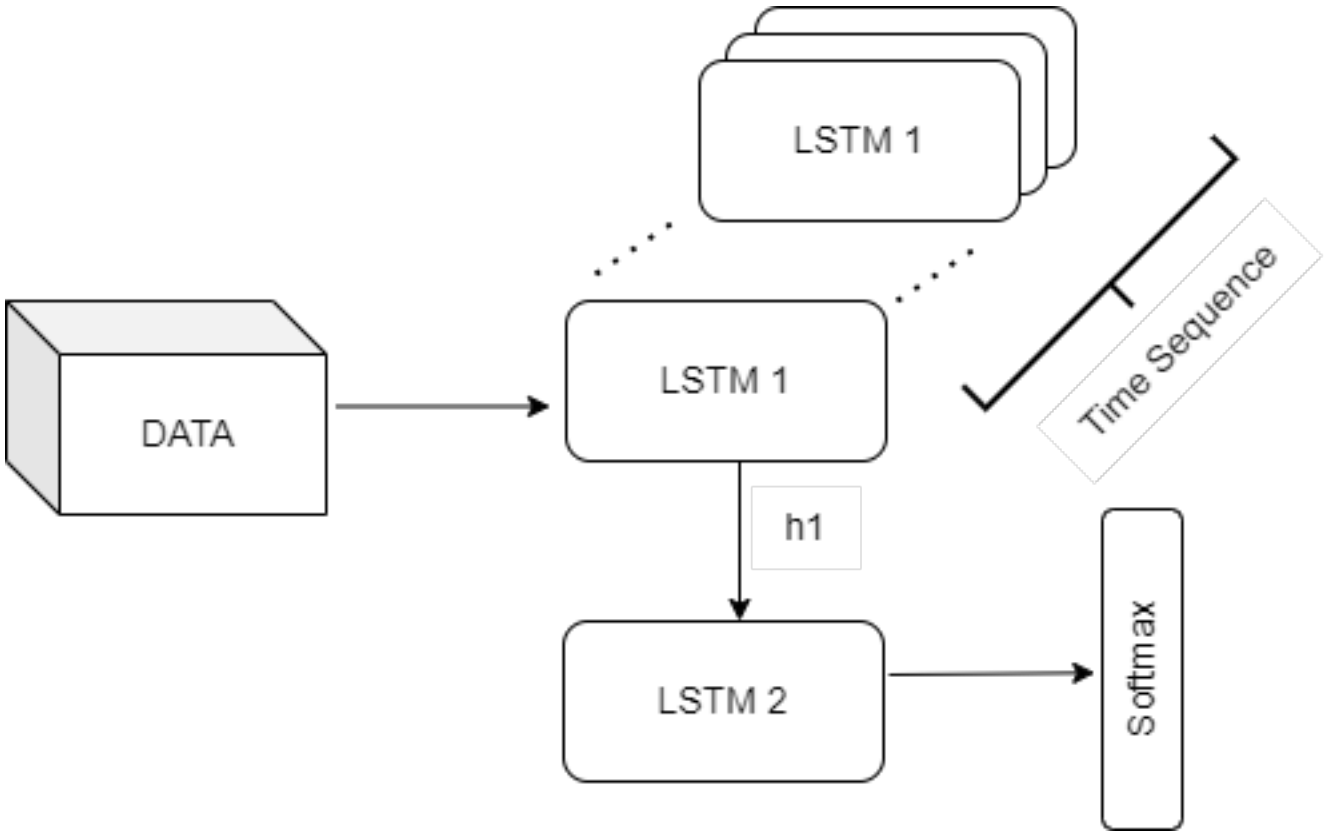

Figure 3.19: Stacked LSTM Architecture
A stacked LSTM is defined as a LSTM model comprising LSTM layers.

No.	Logistic Regression – RFE Top 15 Features
1	BIO Age
2	BIO Marital Status Category
3	BIO Title Category
4	EDU Total UVic
5	EVT Total Registered
6	GEO Phone
7	GEO Total Address count
8	COM Email Received
9	COM Email Opened
10	COM Email Clicked Through
11	GIFT Loyalty Years Donated
12	GIFT First Gift Amount
13	GIFT Total Gift Amount
14	GIFT Age at First Gift
15	GIFT Age at Largest Gift

Table 3.4: Logistic Regression RFE Top 15 Features.

GRU because they are able to approximate non-linear functions in time-dependent sequence data. Also, these models solve the vanishing gradient problem RNN suffers by incorporating three gates, input, output, and memory gate. Recurrent layers of sequential architecture examine temporal correlations of sequence data to learn the time dependencies. [41]. The model architectures are discussed in detail in subsection 3.2. We experimented a stacked sequential architecture (Fiture 3.19). Stacked LSTM is defined as a LSTM model that comprises more than one LSTM layer. By stacking hidden layers, it adds more depth to the model that is considered to be a way of optimizing models. We used stacked architecture for both LSTM and GRU model. The same feature selection methods used for the traditional models were employed on the demographic features for the sequential models.

LSTM/GRU Variants - Bidirectional Model

Bidirectional LSTM (Figure 3.20) is an extended version of the LSTM model where the inputs are fed into two LSTM layers, forward network, and backward network. The output generated from two networks is concatenated and is fed into the softmax classifier for classification. [47]

Bidirectional LSTM feeds inputs both from past to future and from future to past.

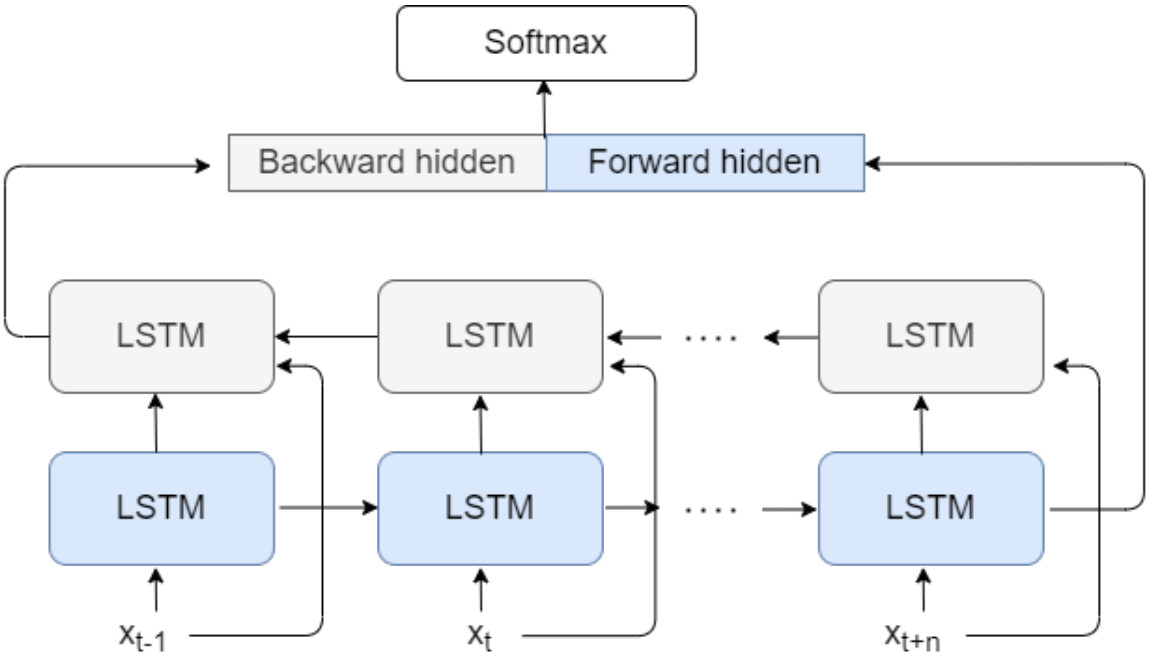

Figure 3.20: Bidirectional LSTM (inspired by the diagram by Victor Makrenkov [47])

It preserves information from the future and combines two hidden states (backward and forward). This allows preserving information from both the past and the future. We used Keras Bidirectional wrapper package for Bidirectional model.

LSTM/GRU Variants - Time distributed Model

The Time-distributed layer (Figure 3-21) is added to the LSTM or GRU architecture to predict the output for each time-step individually. By adding Time-distributed layer, each LSTM/GRU unit returns the output value at each time step instead of a single output as in the regular LSTM/GRU. There are ten values from the ten-year time steps in our model. We used the final year output to predict the future donation amount level for performance evaluation.

Bimodal Model - LSTM/GRU with Condition

Our second goal in this study is to explore a Bimodal model to incorporate both time-variant and time-invariant features into the model. We used LSTM model to process sequences of time-variant features. Wang *et al.*[8] discussed in their literature four options to combine both time-independent and dependent observations in a model.

1. Treat the static features as dynamic data at each stage.

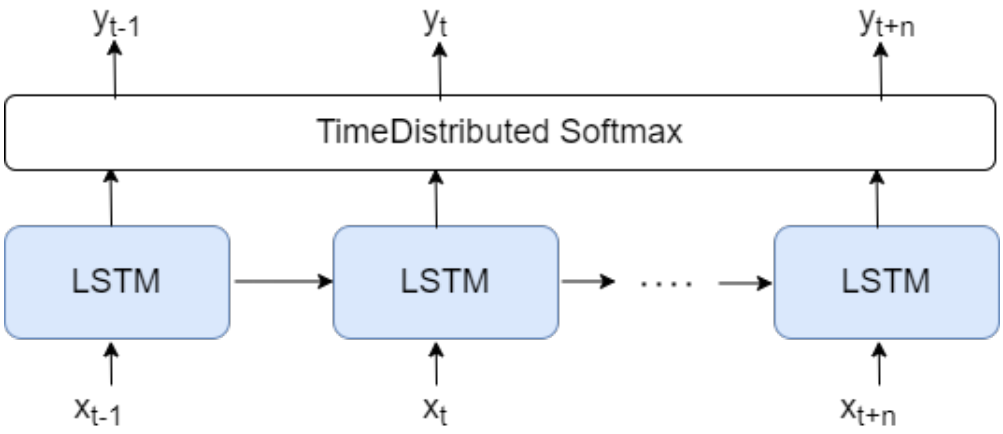

Figure 3.21: Time Distributed LSTM (inspired by the diagram by Dipesh Gautam et al. [18])

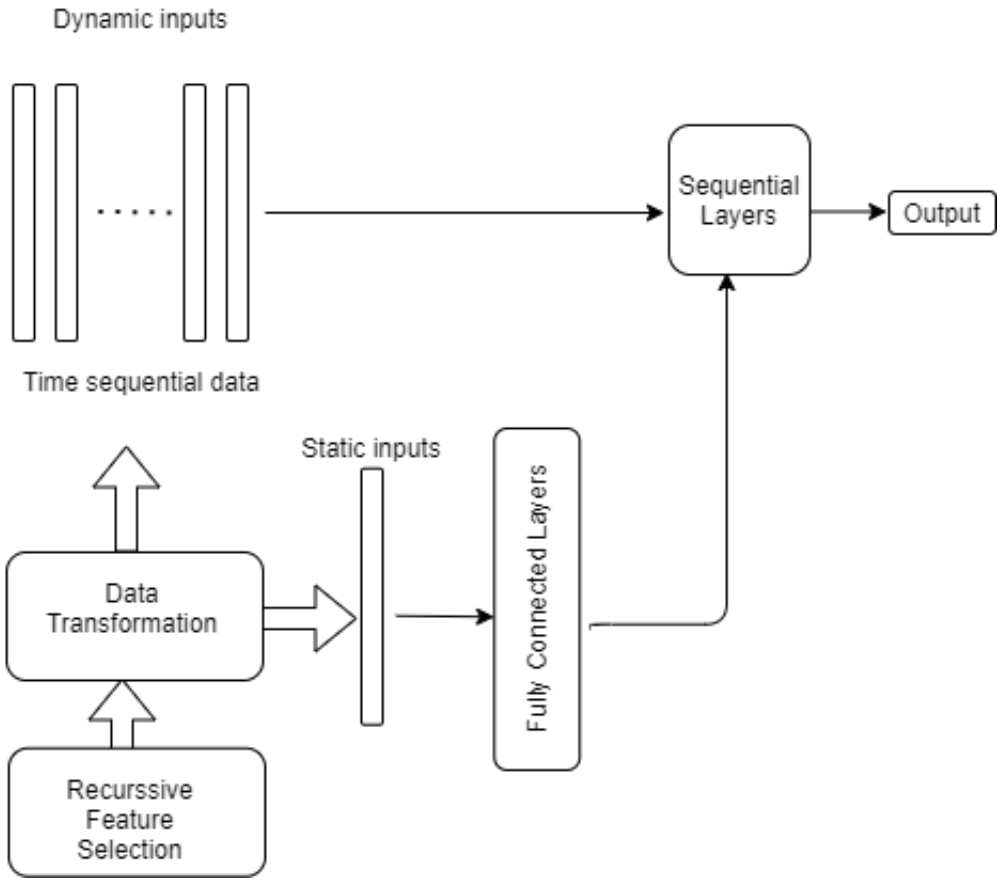

Figure 3.22: Model Architecture and Pipeline
The normalized data selected in the feature selection step is arranged into time-variant and time-invariant features. The time-invariant features are concatenated with the sequential layers outcome and processed together for the last outcome. Sequential Layer is either LSTM or GRU layer.

2. Concatenate the static features with the output of the hidden state at the last time step.

3. Apply the static time-invariant features to initialize the first hidden layer (t = 1) and concatenated with the dynamic features at the following stages (t > 1)

4. Concatenate the static features with the hidden state at each time step.

Each method is discussed in detail in the literature review section, Chapter 2. We adopted the second option among existing methods of combining two types of data. The justification for selecting the second option is time synchronicity of the last time step (t) of sequential learning and the demographic data. For instance, the age factor affects a donor's decision to give. As demonstrated by figures 3.8 to 3.12, there is a correlation between the donor's age and the giving amount.

In the bimodal model (Figure 3.22), the time-variant features include series of annually calculated donation variables, event attendance per year, and open/click-through counts per year. Those features are used to examine whether temporal patterns exist to predict the size of future donations. All time-variant features are summed up annually for the past eleven years including the target year. The time-invariant features are the alumni's age, family information, wealth principal component values based on the area alumni live, relationship with organizations such as UVic boards, family foundations, employers. Those kinds of time-invariant information are treated as if they are conditions to the time-variant features for each alumnus case. The target variable is the annual total amount donation amount shifted by one year from the training time-series data. Based on the distribution of the total annual donation amount, the annual donation amounts were grouped by \$0 (no donation), greater than \$0 and less \$1,000, equal to and greater than \$1,000 and less than \$10,000, and greater than \$10,000. The model is constructed as a multi-class model around these four donation buckets. The accuracy rate, precision, recall, F1 scores, and Cohen's kappa are used to assess model performance. The performance of sequential models, including LSTM and GRU, was compared against the traditional time-invariant models.

Our objectives in this study are as follows: (1) Experiment with sequential learning in the fundraising domain. (2) Extend the sequential model by blending static and dynamic data in the model. (3) Conduct performance comparison and examine the use of sequential learning in the fundraising domain. For the first objective, we

tested LSTM and GRU. The main advantage of LSTM and GRU for the sequences over standard neural networks is that weights are shared across time steps when standard neural network can not remember previous inputs. LSTM and GRU also solve the vanishing gradient issue. If there exists any time dependency in the data, LSTM and GRU are suitable algorithms for time sequence data. Features selected by the RFE feature selection process include demographic features such as age, marital status, and other personal profiles. These demographic features are not utilized in the unimodal sequential models while both static and dynamic data were included in the bimodal model for model comparison.

Model List	
Baseline Model	Proposed Model
Neural Network Support Vector Machine	Stacked LSTM Stacked GRU Bidirectional LSTM Bidirectional GRU Time Distributed LSTM Time Distributed GRU Hybrid model - LSTM with Conditions

Table 3.5: Baseline and Proposed models

Chapter 4

Empirical Validation and Discussion

This section describes the data used for our study, the experiment results, and the major model challenges by comparing the performance of the traditional models and the sequential models over their performance. Section 4.1 gives overview of the data, 4.2 discusses the model evaluation metrics, and 4.3 the test results and takeaway from the study.

4.1 Data Overview

The data we used for this study comes from the University of Victoria, Advancement Services. The dataset consists of 171,874 live constituents in total. Out of those records, 123,515 records are University of Victoria (UVic) alumni. Each record includes personal information such as age, sex, marital status, degree, open email rate, and giving history. For this study, we excluded non-alumni records that often lack important information, such as education data (degree, year of graduation, area of study, etc.) and age. The alumni records includes the forerunner alumni. According to Wikipedia, "the University traces its roots to Victoria College, the first post-secondary institution established in British Columbia in 1903. The provincial normal school was later merged with Victoria College in 1956, becoming the College's Faculty of Education. It operated as an affiliated college until 1963 when it was reorganized into the present University of Victoria." [2] The forerunner alumni are those from Victoria College and Normal School. The alumni donation records trace

back to 1987. Out of 123,515 live alumni, 18,482 alumni are donors, while the rest of alumni, 105,033 are non-donors. The proportion of donors is 15%, which indicates the majority of the UVic alumni are a non-donor. This introduces the class imbalance problem in the classification models, which is addressed in the discussion section 4.2 in detail.

Type	**Total Count**	**%**
Individual live Constituents	171,874	100%
live Alumni	123,515	72%
live Non-Alumni	48359	28%

Table 4.1: live Alumni vs. live Non-Alumni Distribution

Type	**Total Count**	**%**
Live Alumni	123,515	100%
Live Alumni Donors	18,482	15%
Live Alumni Non-Donor	105,033	85%

Table 4.2: Live Alumni Donors vs. Live Alumni Non-Donors

4.1.1 Data Limitations

There are a variety of reasons pertaining to alumni donation to universities. Kosse (2019) [25] researched the correlation between alumni affinity to their university and their financial contribution to the university. He concludes that there was a high correlation between the affinities and gifts in the dataset he used. We also examined the correlation among contributing features that promote alumni's motivation for giving to UVic. However, one of the data challenges for the study of the correlation among feature variables and target variables is missing values; some data types, such as age, employment details, contactable address, and marital status, are missing in many records. The information about the job and marital status are collected by a

self-report or through the conversation with an alumnus. UVic has a well-established student calling program, in which the student callers are trained to solicit alumni for their monetary support. During the phone conversation with alumni, employment, family, and contact information are collected as long as alumni are willing to share; however, the number of questions and types of questions are restricted due to the strict guidelines for privacy and confidentiality of personal information. Additionally, the Advancement Services Office in the University of Victoria has an alumni tracing program to track down the lost alumni. Some types of information are hard to obtain through the calling or alumni tracing program. Therefore, there are still many records lacking values in some features. For the categorical features, the number -1 is assigned to the missing data to indicate that the data is not available. All categorical features are transformed to dummy variables for each category. The missing numerical data, such as age and census data (average income, average real estate value, etc.), was synthesized by Multivariate imputation by chained equations (MICE). The missing data imputation method is discussed in Chapter 3.

Type	**Total Count**	**%**
Live Constituents	171,874	100%
Age Missing	22,455	13%
Age Known	149,419	87%

Table 4.3: Missing Age Among live Constituents

4.1.2 Model Data

As Table 3-4 and 3-5 illustrate, the data suffers from missing data issues. The non-alumni data has a higher percentage of missing data compared to the alumni data, especially in the education field, in addition to the missing data common to both alumni and non-alumni. Because of the higher percentage of missing data in the non-alumni data, the model focuses on alumni data, and the non-alumni data is to be studied separately for future work. As discussed in the previous section, the advantage of Deep Neural Network (DNN) is more salient when the size of the training dataset is large. By reducing the size of training data, therefore, the efficacy of DNN may be affected as discussed.

The UVic alumni data consists of time-invariant and time-variant attributes. The time-invariant data includes demographic information (age, sex, employment, city,

Type	**Total Count**	**%**
live Alumni	123,515	100%
Age Missing	2,601	2%
Age Known	120,914	98%
Work Information Missing	102,679	83%
Work Information Known	20,836	17%
LinkedIn Data Missing	119,218	97%
LinkedIn Data Known	4,297	3%
Live Contactable Alumni	102,794	83%
Living Non-Contactable Alumni	20,721	17%
Marital Status Missing	79,078	64%
Marital Status Known	44,437	36%
Family Information Missing	115,419	93%
Family Information Known	8,096	7%

Table 4.4: Missing Data in Alumni Records

title, and education), census data (average income, average house value, and average debts of the constituent's dissemination area). Those types of information remain the same throughout the fixed period for the model. The time-variant data are donation information for each year for the last ten years starting from 2010. The records before 2010 are legacy data stored in an old format, which is not systematic or coherent. Some information, such as event attendance, campaign, appeals, and family information, is either inaccurate, not collected, or nonexistent. The time-variant gift data has been recorded since 1987, while some time-variant information such as appeals and campaigns have been recorded more consistently since 2010. Therefore, we only used the post-2010 time-variant data for the model.

There are 84 features including categorical dummy variables in the original dataset for feature selection. The feature selection process is run against the largest giving amount to selects the relevant features for the models. The details of the feature selection method and the selected features are discussed in Chapter 3. The multi-class models have four classes:

- Class 1 - $0
- Class 2 - greater than $1 and less than $1,000
- Class 3 - greater than $1,000 to less than $10,000
- Class 4 - greater than $10,000

Type	Total Count	%
Non-Alumni	48,359	100%
Education Missing or Unknown	32,989	68%
Education Known	15,370	32%
Age Missing	33,826	70%
Age Known	14,533	30%
Work Information Missing	48,204	99.7%
Work Information Known	155	0.3%
No LinkedIn Information Available	-	-
Marital Status Missing	38,945	81%
Marital Status Known	9,414	19%
Family Information Missing	38,381	79%
Family Information Known	9,978	21%

Table 4.5: Missing Data Among live Non-alumni

4.2 Evaluation Metrics

Accuracy, recall, precision, F1 scores are the most commonly used evaluation metrics for machine learning performance. Accuracy is the most intuitive evaluation measure, which is simply the ratio of the correctly predicted observations and the total number of observations. The formula of accuracy is (TP+TN)/(TP+TN+FP+FN) where the actual class agrees with the predicted class for TP & TN while the actual class contradicts with the predicted class for FP & FN.

TP = True positive
TN = True negative
FP = False positive
FN = False negative

Accuracy is an common metric for model performance evaluation; however, it does not always disclose the full picture. The problems with accuracy include:

- Accuracy does not distinguish between the types of errors it makes, that is False Positive and False Negatives [23].
- Accuracy is a good measure only when the dataset is symmetric, where the counts of false positive and false negative are almost the same (balanced data).

Precision is the ratio of correctly predicted positive observation (TP) to the total

predicted positive observation (TP+FP), while recall is the ratio of correctly predicted positive observations (TP) to all actual positive observations (TP+FN). Lastly, F1 score is the weighted average of precision and recall. We adapted these four evaluation indicators to compare the traditional and our bimodal models.

Accuracy = (TP+TN)/(TP+FP+TN+FN)
Precision = TP/(TP+FP)
Recall = TP/(TP+FN)
F1 Score = 2 * (Recall * Precision) /(Recall + Precision)

The models experimented with:

- Time-invariant Multiclass Neural Network Model
- Time-invariant Multiclass Neural Network with SMOTE
- Time-invariant Multiclass Support Vector Machine
- Time-variant Multiclass Stacked LSTM
- Time-variant Multiclass Stacked GRU
- Time-variant Multiclass LSTM Bidirectional
- Time-variant Multiclass GRU Bidirectional
- Time-variant Multiclass LSTM TimeDistributed
- Time-variant Multiclass GRU TimeDistributed
- Time-variant Multiclass LSTM with Time-invariant Conditions (bimodal Model)
- Time-variant Multiclass GRU with Time-Invariant Conditions (bimodal Model)

NOTE: SMOTE stands for Synthetic Minority Oversampling Technique.

Tables 4.2 to 4.5 are the confusion metrics for the selected models out of all experiments, and table 4.6 describes the total recall rate combining class 1, 2 and 3 recall rate, which is the total numbers of correctly predicted classed out of all minority classes. The overall recall rate is correctly predicted giving greater than zero out of true 1, 2 and 3.

Figure 4.1 - 4.4 are the diagnostic performance charts for the selected sequential models. They are used to determine the number of epoch for each sequential model. The number of epochs is one of the hyper-parameters that defines how many times the learning algorithm runs through the training dataset where each epoch updates the model parameters. Figures 4.1 and 4.2 show that the validation loss starts increasing around epoch 100. The validation loss of time Distributed model drops sharply at the early stage. The training keeps improving gradually while validation loss stays the same, as shown in Figure 4.3. The LSTM with conditions model is not learning after the immediate drop of the validation loss, as shown in Figure 4.4. Several epoch numbers are examined to select the optimal epoch number for each model.

4.3 Results

This study aims to study new approaches as well as to conduct a comparative analysis of traditional and bimodal sequential models in fundraising. We also extend the sequen- tial method by concatenating sequential features and time-invariant features in the bimodal architecture. We evaluates these model performance, by reviewing accuracy, recall, and Cohen's kappa for comparison in this section.

4.3.1 Cohen's Kappa for Class Imbalance

Anand *et al.* [2], in their study, demonstrated the majority class dominates the net gradient that is accountable for updating the model's weights. It indicates that the error of the majority class decreases very fast at an early iteration of the gradient descent. In contrast, the error of the minority group increases and converges slowly. We used Cohen's Kappa score to evaluate the robustness of the models for the imbalanced data. Folorunso [19] discussed the use of the Cohen's Kappa when the dataset suffers class imbalance. It is used to measure the agreement between predicted and observed categorized values. Cohen's kappa ranges from -1 to 1 and has a more realistic view of the model performance for the dataset with severe class imbalance. McHugh (2012) suggests the general guideline for the score as follows.

- less & equal to 0 - no agreement
- 0.01–0.20 : none to slight
- 0.21–0.40 : fair

- 0.41–0.60 : moderate
- 0.61-0.80 : substantial
- 0.81–1.00 : almost perfect agreement

One of the solutions to mitigate class imbalance issue is SMOTE. We applied SMOTE to both traditional and sequential models in our study. As table 4.1 shows, the NN model in traditional architecture achieved high accuracy and recall rate. The result is equivalent after application of SMOTE. The sequential model with SMOTE, on the other hand, behaved highly erratically and often failed to converge; therefore, both Neural Netowrk with SMOTE and bimodal model with SMOTE were omitted in the performance comparison.

The datasets used for all models, including Neural Network, have an uneven distribution of observations in target observations. Under the same condition, the Neural Network model achieved 61% (substantial) of Kappa score without re-balancing class distribution (Tale 4.7). SVM, Bidirectional GRU, Time Distributed LSTM scored 0.21-0.40 (fair) range. The rest of the models are in the 0.01-0.20 (no or slight) range. The results indicate that Neural Networks are the best performer for the imbalanced class dataset used for our study. The structure of the SVM model is the same as in the neural network model, yet its performance is significantly different. It implies that the traditional model does not necessarily outperform the sequential models.

4.3.2 Recall Rate

In the fundraising context, false negative results in lost revenue and the opportunity for financial support in the fundraising model. On the other hand, false positive is interpreted as the loss of human and financial resources allocated to solicit false positives. A false negative is more costly in fundraising than a false positive; therefore, our study focuses on the recall rates for performance comparison. As shown in Table 4.6, Neural Network achieved 98% for majority class, 91% for minority class 1, 72% for minority class 2, and 0% for minority class 3. The recall rates of the traditional SVM are almost equivalent to or lower than some sequential models. Again, these results indicate that we can not conclude that the traditional models outperform the sequential models. Instead, we could deduce the conclusion that Neural Network has the ability to extract meanings from the data and predict the future giving amount range. Even the high performing neural network model could not identify any of

the class 3 observations. Given that observations from a minority class consist of a negligible 0.4% of all observations, the class 3 recall rates, 20 - 28% of the SVM and Time Distributed models, are considered good.

We also evaluated model performance on the selected models by focusing on minority recall rates. Two rates, minority recall and overall (binary) recall, were calculated based on the confusion matrix. The minority recall rate is the percentage of minority true positives (class 1, 2, & 3) divided by all minority observations. The overall recall rate is the percentage of all predicted minority observations, regardless of their classes, divided by all minority observations (Table 4.3). The stacked LSTM model identified 63.2% observations donating in 2020 out of all minority classes. It is better than random guessing, which should be equal to the proportion of the minority group, 23.5%. Neural Networks achieved 89.2% for minority recall and a remarkable 99.7% overall recall rate. This recall rate assessment proves that neural networks is a most robust and reliable model.

4.3.3 Limited Sequential Data and Performance

This section discusses the sample size challenge with the sequential models. Cui *et al.* [12] state in their study that neural networks could fail to generalize from the training data when training data is not sufficient. They also indicate that one needs careful consideration of sample size issues to achieve a successful network generalization. The alumni dataset contains enough observations; however, the time sequence data is limited due to the data consistency issues discussed in Chapter 3. The feature data such as event registrations, click-through and appeals are nonexistent or not consistent before 2010, consequently limiting the number of time-variant sequences. The time-variant window, therefore, is limited to the period from 2010 to 2020. Thus it could intensify the class imbalance in each time step because not all donors give financial support every year. The short time sequences, in our study ten time steps, do not give enough context for an accurate prediction, especially when the data suffers sever class imbalance. Each alumnus observations include number of features times 10 sequences, meaning 10 observations per feature. Despite sufficient observations for each time step, the limited number of time step hampers the sequential model from learning temporal patterns. Sequential neural networks shares architectural similarities with simple neural networks. The performance dropped significantly by adding time sequences to the neural networks. It is indicative of a need for further studies

on performance gain by increasing sequential data. The possible solutions to this problem could be:

1. to remove other time-variant feature and only use gift-related features, which have consistently collected since 1987

2. to increase the frequency of measures such as quarterly or monthly to increase the number of temporal sequences.

4.3.4 Bimodal Model and Performance

Bimodal sequential model combines sequential learning and a fully connected static layer within the model. The model failed to present a high accuracy that generalizes the trained model to classify unseen data. Although minority class is harder to predict for the sequential neural networks because of few examples of each minority class at each time step, the other sequential models achieved a high accuracy rate in the 92 - 98% range. On the other hand, the bimodal model could not reach as high accuracy as other models. The bimodal model differs from the other sequential models for its static conditions. In other words, concatenating the static data affected adversely. We could infer from these results the following. First, the LSTM output data does not correlate well with the other information added and adds white noise to the model. Second, it requires more training on the blended data (LSTM time sequence data and static data) to benefit from the additional static data. Two options are suggested for future study.

- fully connected neural networks (FC NN) to be employed on the concatenated data (LSTM output and static data)

- to run the fully connected neural network (FC NN) on the time-invariant features and concatenate the output with the sequential data before running LSTM.

- to run the fully connected neural network (FC NN) on the time-invariant features separately and concatenate the output with the sequential output after running LSTM.

Comparison Summary

	Traditional Models	LSTM/GRU	LSTM w/ Conditions
Accuracy	High	High	Lower than random
Recall Minority (Each class)	High - NN Low - SVM	Low for most of minority classes	Low in all minority classes
Recall Minority (Overall)	High - NN Low - SVM	Stacked LSTM 68.62%	Low
Use	NN for binary/multiclass	Possibly used for binary models	Experiment on running another model after concatenating static data

Figure 4.1: Model Comparison Summary Chart

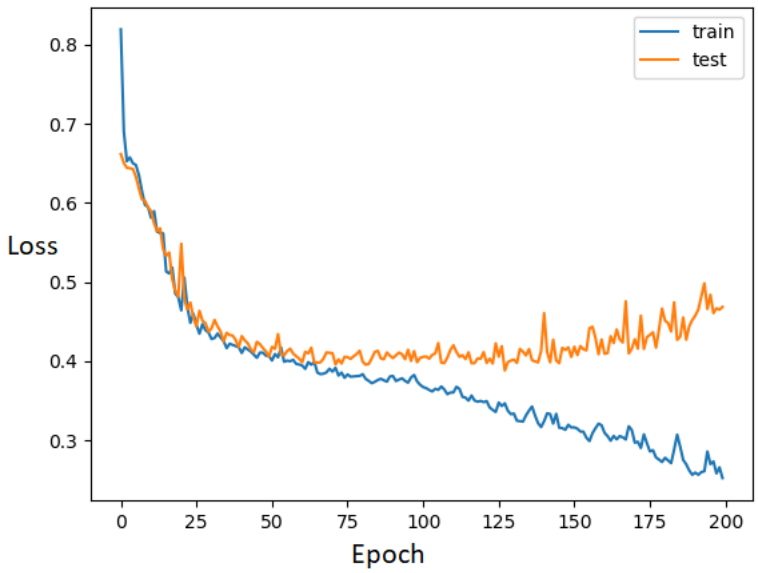

Figure 4.2: Validation Loss Chart - Stacked LSTM

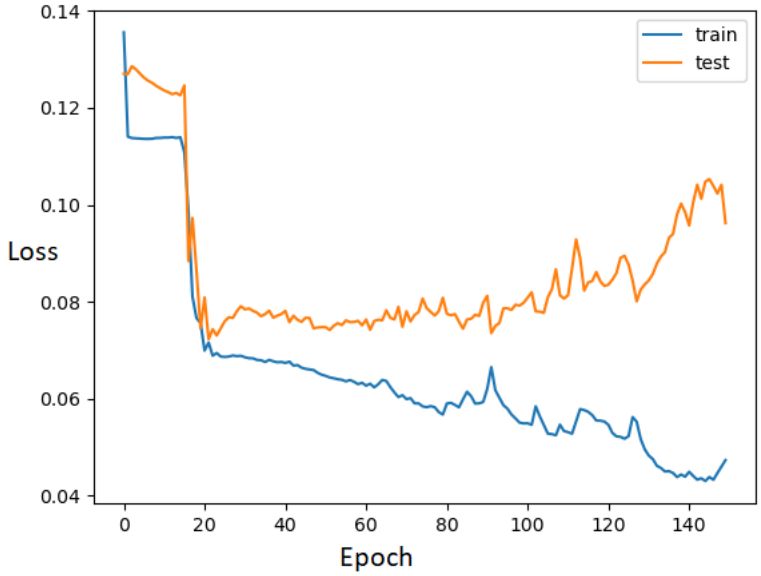

Figure 4.3: Validation Loss Chart - Stacked GRU

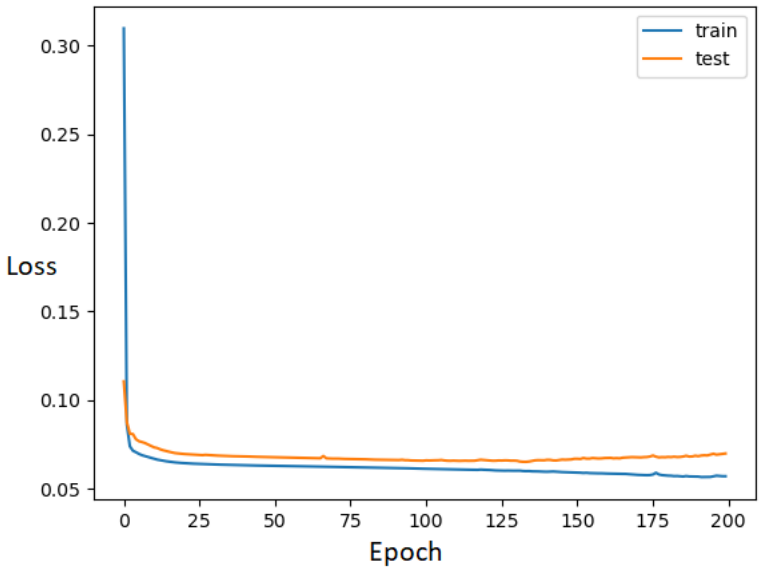

Figure 4.4: Validation Loss Chart - Time Distributed LSTM

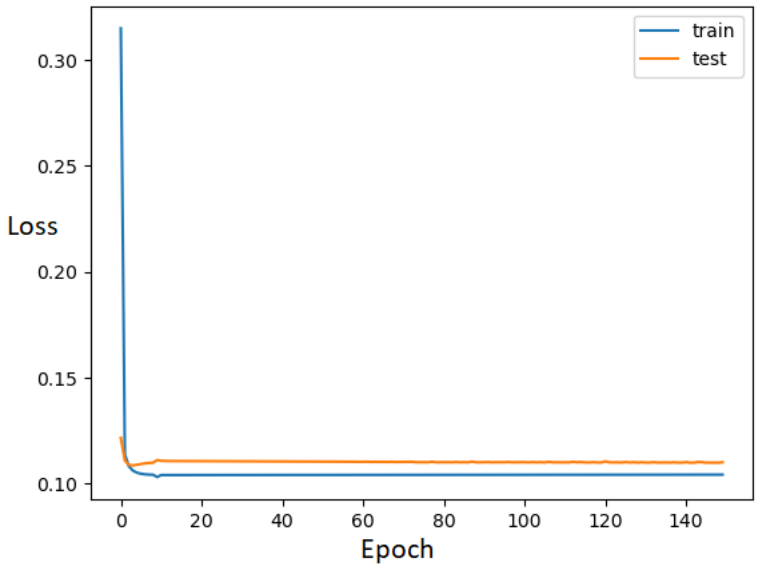

Figure 4.5: Validation Loss Chart - LSTM with Conditions

Model	Class	Precision	Recall	F1	Accuracy
NN	0	1.00	0.98	0.91	0.98
	1	0.48	0.91	0.63	0.98
	2	0.20	0.72	0.32	0.98
	3	0.00	0.00	0.00	0.98
SVM	0	0.85	0.97	0.90	0.83
	1	0.60	0.23	0.34	0.83
	2	0.22	0.05	0.08	0.83
	3	0.25	0.27	0.26	0.83
Stacked LSTM	0	0.99	0.94	0.97	0.92
	1	0.10	0.13	0.12	0.92
	2	0.02	0.75	0.12	0.92
	3	0.00	0.00	0.00	0.92
Stacked GRU	0	0.99	0.95	0.97	0.92
	1	0.21	0.15	0.18	0.92
	2	0.02	0.55	0.3	0.92
	3	0.00	0.00	0.00	0.92
LSTM Bidirectional	0	0.99	0.88	0.93	0.88
	1	0.12	0.53	0.19	0.88
	2	0.01	0.02	0.02	0.88
	3	0.00	0.00	0.00	0.88
GRU Bidirectional	0	0.99	0.98	0.98	0.96
	1	0.21	0.28	0.24	0.96
	2	0.08	0.18	0.11	0.96
	3	0.00	0.00	0.00	0.96
LSTM TimeDistributed	0	0.99	0.99	0.99	0.98
	1	0.48	0.43	0.45	0.98
	2	0.01	0.02	0.02	0.98
	3	0.08	0.22	0.12	0.98
GRU TimeDistributed	0	0.99	0.98	0.98	0.97
	1	0.41	0.30	0.34	0.97
	2	0.10	0.45	0.17	0.97
	3	0.01	0.22	0.01	0.97
LSTM with Conditions	0	0.98	0.72	0.83	0.71
	1	0.02	0.28	0.04	0.71
	2	0.00	0.00	0.00	0.71
	3	0.00	0.00	0.00	0.71

Table 4.6: Experiment Results (NOTE: GRU with Conditions is omitted due to the failure of minority identification.)

Model	Kappa
Neural Network	0.612760
SVM	0.282066
Stacked LSTM	0.167466
Stacked GRU	0.184376
Bidirectional LSTM	0.143740
Bidirectional GRU	0.238634
Time Distributed LSTM	0.280993
Time Distributed GRU	0.205246

Table 4.7: Cohen's Kappa Score

	Predict 0	**Predict 1**	**Predict 2**	**Predict 3**
True 0	34484	667	69	23
True 1	2	621	47	9
True 2	0	9	31	3
True 3	0	3	6	0

Table 4.8: Confusion Matrix Neural Network with no time-variant features

	Predict 0	**Predict 1**	**Predict 2**	**Predict 3**
True 0	32436	809	1260	1
True 1	258	83	323	0
True 2	4	6	32	0
True 3	1	3	5	0

Table 4.9: Confusion Matrix Stacked LSTM

	Predict 0	**Predict 1**	**Predict 2**	**Predict 3**
True 0	34133	285	70	18
True 1	376	283	4	1
True 2	14	23	1	4
True 3	5	2	0	2

Table 4.10: Confusion Matrix LSTM Time Distributed Model

	Predict 0	**Predict 1**	**Predict 2**	**Predict 3**
True 0	33824	289	112	281
True 1	388	198	52	26
True 2	15	1	19	9
True 3	5	0	2	2

Table 4.11: Confusion Matrix GRU Time Distributed Model

Model	**Class 0 Recall**	**Minority Recall**	**Minority Overall**
NN	98.0%	89.2%	99.7%
Stacked LSTM	94.0%	16.1%	63.2%
LSTM TimeDistributed	99.0%	40.0%	44.8%
GRU TimeDistributed	98.0%	30.5%	43.2%

Table 4.12: Minority Recall Rate - Selected Models

Chapter 5

Conclusions

While there are many machine learning applications in various fields, the use of machine learning to optimize non-profit fundraising activities is still limited. This study introduces a deep sequential learning algorithm that few or no researchers have used before in fundraising and compares its performance against traditional models.

The traditional approach to fundraising optimization problems relies on the data collected at one specific point in time. However, the real-world fundraising data contains not only time-invariant data but also time-varying data. To the best of our knowledge, there is no research on applying a sequential model in the fundraising domain. In our study, we experimented with using time-varying data in the model. Additionally, we studied the hybrid approach by introducing bimodality in the model that includes both time-variant and time-invariant features in the models.

Among all the models tested, traditional unimodal neural networks demonstrated the most capable of achieving strong classification performance. The performance of SVM in the traditional structure is almost equivalent to the sequential model performance in every evaluation metric. By comparing these models, we identified that the traditional models do not always outperform the sequential models. To put it another way, it depends on the architecture. The unimodal neural networks proved most capable of learning giving patterns and are the most robust and reliable model on highly imbalanced alumni data.

There are a few possible causes of lower recall and Cohen's kappa rates among the sequential models; first, limited time sequence, and second, magnified class imbalance at each time step. Due to the database quality issue, the sequential data was collected from a shorter period, 2010 to 2020. This introduces the sample size issue of a short time steps. Sequential neural networks could fail to generalize from the training data

when training time sequence data is not sufficient. Secondly, the class imbalance is magnified when the model looks at each time step while the traditional model looks at overall giving data. Covering longer time sequence could be a solution.

The cause of the poor performance of the bimodal model is thought to be the process after concatenating the LSTM layer and static data. The Hybrid model evaluation metrics are lower than those of the other sequential models. It could be because the LSTM data does not correlate well with the other information added, meaning adding noise to the model, or it requires more training on the blended data (LSTM time sequence data and static data) to benefit from the additional static data. There are a couple of methods learning from the static data by using fully connected neural networks discussed in Chapter 4. More research is needed to confirm whether running additional learning on blended data helps improve its performance.

The sequential models achieved high accuracy, which is much better than random guessing. Despite the low recall rate for each class, the overall binary recall rate indicates its potential to identify time dependent patterns. The empirical results indicate that the sequential model could be used for binary models.

Donor retention is an essential factor for successful fundraising. The churn model is used to predict at-risk donors and take proactive measures to prevent the existing donors from leaving. The predicted probability that a donor will churn is used in designing donor retention strategies. The results from sequential learning may provide inspiration for developing proactive strategy for at-risk donors and reducing the risk of donor churning.

Another potential area we could use the sequential model is communication. The University of Victoria has sufficient fundraiser auction records with alumni, including action type (phone call, mail, meeting), frequency, etc. These types of action records also suffer data limitations in time sequence but cover a longer time window than other types of sequential data. Wang *et al.* [8] used time-variant communication flow data to predict the success of crowdfunding. Their study proved that the communication flow data is critical to predicting the outcome in a timely manner. The use of action data in the fundraising model potentially provides the model success.
The results demonstrate the potential use of neural networks and a unimodal sequential model in the prediction of fundraising outcomes.

Lastly, if we were to design this study again, there are a number of changes or experiment we would make.

- Rebalancing time-variant data
  SMOTE application in the time-variant model caused erratic behavior and failed to converge; however, there are other synthesizing methods, including autoencoder, and hybrid method that combines over-sampling and under-sampling (SMOTE ENN and SMOTE Tomek Link) [5].

- Cost-sensitive learning
  Cost-sensitive learning introduces a misclassification cost to minimize the conditional risk, i.e., strongly penalizing misclassification of the minority classes [6].

- Running fully-connected neural network (FC NN) before LSTM
  Run the fully connected neural network on the time-invariant features and concatenate the output with the sequential model data. The concatenated data is used for LSTM sequential learning.

- Running fully-connected neural network (FC NN) after LSTM
  Run the fully connected neural network on the concatenated data (sequential output and time-variant data.

- Running fully-connected neural network (FC NN) and LSTM separately
  Run the fully connected neural network on the time-invariant features separetely and concatenate the output with the sequential model output after running LSTM. Then the concatenate two outputs is used for LSTM sequential learning.

- Various hyper-parameter optimization methods
  We applied grid search method on the traditional SVM model; however, grid search is too expensive and it was not practical to apply it on the other models. Instead of grid search, alternative options could be applied and compared for their performance.
  Widely used optimization methods [51]:

  Manual search

  Grid search

  Random search

  Bayesian optimization

  Gradient-based optimization

Bibliography

[1] Jinhu Bian Guangbin Lei Meisam Amani Amin Naboureh, Ainong Li. *A Hybrid Data Balancing Method for Classification of Imbalanced Training Data within Google Earth Engine: Case Studies from Mountainous Regions.* Remote Sensing 12.20 (2020): 3301. Print., 2020.

[2] R. Anand, K.G. Mehrotra, C.K. Mohan, and S. Ranka. *An improved algorithm for neural network classification of imbalanced training sets.* IEEE Transactions on Neural Networks, Volume 4, No. 6, pp. 962-969, doi: 10.1109/72.286891, 1993.

[3] Yurekli A. Bilge A. Kaleli C. Batmaz, Z. *A review on deep learning for recommender systems: challenges and remedies.* Artificial Intelligence Review, 52(1): 1-37, 2019.

[4] Yang Jiang, N. Bosch, R. Baker, L. Paquette, J. Ocumpaugh, J. M. A. L. Andres, A. Moore, Gautam Biswas. Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection? In *AIED*, 2018.

[5] Labrador Mirador Yang Song Weiyan Hou Chao Liu, Jia Wu. *Classifying DNA Methylation Imbalance Data in Cancer Risk Prediction Using SMOTE and Tomek Link Methods.* Data Science, 2018, Volume 902 ISBN : 978-981-13-2205-1, 2018.

[6] Victor S. Sheng Charles X. Ling. *Cost-Sensitive Learning and the Class Imbalance Problem.* Encyclopedia of Machine Learning. C. Sammut (Ed.). Springer. 2008, 2008.

[7] Quin Chen. *Predictive Modeling For Non-profit Fundraising.* James Madison University, 2010.

[8] T. Wang, F. Jin, Y. Hu, Y. Cheng. *Early Predictions for Medical Crowdfunding: A Deep Learning Approach Using Diverse Inputs.* ArXiv, abs/1911.05702, 2019.

[9] Kyoungnam Ha, Sungzoon Cho, and Douglas Maclachlan. *Response Model on Bagging Neural Networks*. Journal of Interactive Marketing Volume 19/Number 1/Winter 2005, 2005.

[10] Guozheng Rao, Weihang Huang, Zhiyong Feng, Qiong Cong. *LSTM with sentence representations for document-level sentiment classification*. Neurocomputing, Volume 308, 2018.

[11] Stanford University CS231. Is it time to swish? comparing deep learning activation functions across nlp tasks. `https://cs231n.github.io/linear-classify/#softmax`.

[12] Ying-Jin Cui, S. Davis, Chao-Kun Cheng, and Xue Bai. *Linking Marketing to Nonprofit Performance*. Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826) vol 6, 2004.

[13] R. Qing dao-er-ji, Y. L. Su, and W. W. Liu. *Research on the LSTM Mongolian and Chinese machine translation based on morpheme encoding*. Neural Computing Applications, 32(1), 2020.

[14] Shuangyin Liu Daoliang Li. `https://www.sciencedirect.com/science/article/pii/B9780128113301000041`.

[15] T. K. Das. *A customer classification prediction model based on machine learning techniques*. International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Davangere, 2015.

[16] Morris George Denish Shah. *Linking Marketing to Nonprofit Performance*. Journal of Public Policy and Marketing, 2021.

[17] Larry H. Dietz. *Iowa State University Alumni Contributions: An Analysis Of Alumni Giving Patterns By Selected Class Years, 1974 And 1979*. Iowa State University, 1985.

[18] Rajendra Banjade Lasang Jimba Tamang Dipesh Gautam, Nabin Maharjan. *Long Short Term Memory based Models for Negation Handling in Tutorial Dialogues*. Negation Handling in Tutorial Dialogues. 10.13140/RG.2.2.26250.36804., 2020.

[19] S.O. Folorunso. *Alleviating Classification Problem of Imbalanced Dataset.* Mathematical Sciences Department, Olabisi Onabanjo University, 2013.

[20] Chris Howard and Andy Rowsell-Jones. *2019 CIO Survey: CIOs Have Awoken to the Importance of AI.* Gartner Inc, 2019.

[21] Angela M. Woodc Ian R. White, Patrick Roystonb. *Multiple imputation using chained equations: Issues and guidance for practice.* Wiley Online Library, 2010.

[22] A.I. in Advancement Advisory Council. The state of ai in advancement report. `https://gravyty.s3.amazonaws.com/2019aaacstateofaiinadvancement.pdf`, 2018.

[23] Nathalie Japkowicz. *Why Question Machine Learning Evaluation Methods? (An illustrative review of the shortcomings of current methods).* School of Information Technology and Engineering University of Ottawa, 2006.

[24] Jennifer Key. *Enhancing Fundraising Success with Customer Data Modelling.* International Journal of Nonprofit and Voluntary Sector Marketing, 2006.

[25] Glenn Kosse. *The Relationship Between Young Alumni Participation and Giving.* Bellarmine University Graduate Theses, Dissertations, and Capstones. 66., 2019.

[26] Philip Kotler. *Strategies for Introducing Marketing into Nonprofit Organizations.* Journal of Marketing, 43 (I), 37-44, 1979.

[27] Caglar Gulcehre Dzmitry Bahdanau Fethi Bougares Holger Schwenk Yoshua Bengio Kyunghyun Cho, Bart van Merrienboer. *Learning Phase Representations Using RNN Encoder-Decoder for Statistical Machine Translation.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). doi:10.3115/v1/d14-1179, 2014.

[28] Duyu Tang, Bing Qin, Ting Liu. *Document Modeling with Gated Recurrent Neural Network for Sentiment Classification.* Harbin Institute of Technology, Harbin, China, 2015.

[29] Dooman Arefan Shandong Wu Long Gao, Lu Yang. *One-class classification for highly imbalanced medical image data.* Proc. SPIE 11318, Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, 113181C (2 March 2020); doi: 10.1117/12.2551389, 2006.

[30] Maciej A. Mazurowski Mateusz Buda1, Atsuto Maki. *A systematic study of the class imbalance problem in convolutional neural networks.* Neural Networks, vol. 106, 2018, pp. 249–259., doi:10.1016/j.neunet.2018.07.011, 2018.

[31] K. Muralidharan. *A Note on Transformation, Standardization and Normalization.* Department of Statistics, Maharaja Sayajirao University of Baroda, 2010.

[32] Washington DC National Council of Nonprofit. `https://www.councilofnonprofits.org/tools-resources/crowdfunding-nonprofits#:~:text=Crowdfunding%20is%20a%20term%20that,ventures%2C%20primarily%20via%20the%20internet`.

[33] C. Olah. `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`.

[34] AR Andreasen P. Kotler. *Strategic Marketing for Nonprofit Organizations.* (International ed.) 6th ed., Pearson Education, New York, NY, 2003.

[35] Ajiboye Abdulraheem, Ruzaini Abdullah Arshah, Hongwu Qin. *Evaluating the Effect of Dataset Size on Predictive Model Using Supervised Learning Technique*, volume 1. 205.

[36] Alex Graves, Abdel rahman Mohamed and Geoffrey Hinton. *Speech Recognition With Deep Recurrent Neural.* Department of Computer Science, University of Toronto, 2013.

[37] Eric Rothstein Morrish Ralf C. Staudemeyer. *Understanding LSTM - a tutorial into Long Short-Term Memory Recurrent Neural Networks.* Schmalkalden University of Applied Sciences, Germany, Singapore University of Technology and Design, Singaporet, 2019.

[38] Tobias Lang Matthias Rettenmeier. *Understanding Consumer Behavior with Recurrent Neural Networks.* 2017.

[39] Hojjat Salehinejad and Shahryar Rahnamayan. *Customer Shopping Pattern Prediction: A Recurrent Neural Network Approach.* arXiv:1804.07669, 2016.

[40] Ajith Abraham Shaza M. Abd Elrahman1. *A Review of Class Imbalance Problem.* Journal of Network and Innovative Computing ISSN 2160-2174, Volume 1 (2013) pp. 332-340, 2013.

[41] Tiberiu Cocias Gigel Macesanu Sorin Grigorescu, Bogdan Trasnea. *A survey of deep learning techniques for autonomous driving.* Journal of Field Robotics, doi: 10.1002/rob.21918, 2020.

[42] A. Srinivasan. *An Ensemble Deep Learning Approach to Explore the Impact of Enticement, Engagement and Experience in Reward Based Crowdfundin.* Department of Computer Science and Engineering SRM Institute of Science and Technology, 2020.

[43] Iryna Gurevych Steffen Eger, Paul Youssef. *Recurrent Neural Networks.* Ubiquitous Knowledge Processing Lab (UKP-TUDA) Department of Computer Science Technische Universitat Darmstadt, 2019.

[44] Luchen Liu, Jianhao Shen, Ming Zhang, Zichang Wang, Jian Tang. *Learning the Joint Representation of Heterogeneous Temporal Events for Clinical Endpoint Prediction.* 2018.

[45] Yanwei Cui, Rogatien Tobossi, and Olivia Vigouroux. *Modelling customer online behaviours with neural networks: applications to conversion prediction and advertising retargeting.* Department of Electrical, Computer, and Software Engineering University of Ontario Institute of Technology, 2018.

[46] Li Z, Tam V. *Combining the real-time wavelet denoising and long- short-term-memory neural network for predicting stock indexes.*

[47] Bracha Shapira Victor Makarenkov, Lior Rokach. *Choosing the Right Word: Using Bidirectional LSTM Tagger for Writing Support Systems.* Department of Software and Information Systems Engineering Ben-Gurion University of the Negev Beer-Sheva, Israel, 2019.

[48] Mark Walcott. *Predictive Modeling and Alumni Fundraising in Higher Education.* Illinois State University, 2014.

[49] Tingwu Wang. *Recurrent Neural Networks.* Machine Learning Group, University of Toronto, FOR CSC 2541, SPORT ANALYTICS.

[50] Chaoran Cheng, Fei Tan, Xiurui Houand, Zhi Wei. *Success Prediction on Crowdfunding with Multimodal Deep Learning.* 2018.

[51] Wikipedia. Crowdfunding. `https://en.wikipedia.org/wiki/Crowdfunding`.

[52] Lian Yan, D. J. Miller, M. C. Mozer, R. Wolniewicz. *Improving prediction of customer behavior in non-stationary environments.* IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222), Washington, DC, USA, 2001.

[53] Zhi-hua Zhou Xu-Ying Liu. *The Influence of Class Imbalance on Cost-Sensitive Learning: An Empirical Study.* Sixth International Conference on Data Mining (ICDM'06) (2006). Print., 2006.

[54] Giha Lee 1 Xuan-Hien Le, Hung Viet Ho and Sungho Jung. *Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting.* Department of Disaster Prevention and Environmental Engineering, Kyungpook National University, Faculty of Water Resources Engineering, Thuyloi University, 2019.

[55] Yong Yu. *A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures.* Department of Automation, Xi'an Institute of High-Technology, 2019.