STATISTICAL AND MACHINE LEARNING METHODS OF ONLINE BEHAVIOURS ANALYSIS

Judah Soobramoney

January, 2024

Statistical and Machine Learning Methods of Online Behaviours Analysis

by

Judah Soobramoney

A thesis submitted to the

University of KwaZulu-Natal

in fulfilment of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

APPLIED STATISTICS

Thesis Supervisor: Prof Retius Chifurira

Thesis Co-supervisor: Prof Temesgen Zewotir

UNIVERSITY OF KWAZULU-NATAL

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE

WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Declaration - Plagiarism

I, Judah Soobramoney, declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowlegded as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then
   (a) their words have been re-written but the general information attributed to them has been referenced, or
   (b) where their exact words have been used, then their writing has been placed in italics and referenced.
5. This thesis does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and in the reference sections.

Judah Soobramoney (Student)	2024/06/06 Date
 Prof Retius Chifurira (Supervisor)	2024/07/27 Date
 Prof Temesgen Zewotir (Co-supervisor)	2024/07/27 Date

Disclaimer

This document describes work undertaken as a PhD programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

ABSTRACT

The success of corporates is highly influenced by the effectiveness and appeal of each corporate's website. This study was conducted on TEKmation, a South African corporate, whose board of directors lacked insight regarding the website's usage. The study aimed to quantify the web-traffic flow, detect the underlying browsing patterns, and validate the web-design effectiveness. The website experienced 7,935 visits and 57,154 pageviews from 1 June 2021 to 30 June 2023 (data sourced by Google Analytics). Grubb's test has identified outliers in visit frequency, the pageviews per visit, and the visit duration per visit. A small degree of missingness was observed on the mobile device branding (1.24%) and operating system (0.03%) features which were imputed using a Bayesian network model. To address a data-shift detected, an artificial neural network (ANN) was proposed to flag future data-shifts with important predictors being the period of year and volume of sessions. Prior to clustering, feature selection methods assessed the feature variability and feature association. Results indicated that low-incidence webpages and features with natural relationships should be omitted. The K-means, DBScan and hierarchical unsupervised machine learning methods were employed to identify the visit personas, labelled *get-in-touch* (12%), *accidentals* (11%), *drop-offs* (30%), *engrossed* (38%) and *seekers* (9%). It was evident that the premature drop-offs needed further exploration. The Cox proportional hazards survival model and the random survival forest (RSF) model have identified that the web browser, visit frequency, device category, distance, certain webpages, volume of hits, and organic searches proved to be drop-offs hazards. A tiered Markov chain model was developed to compute the transition probabilities of dropping-off. The *contact* (63%) and *clients* (50%) states recorded a high likelihood to drop-off early within the visit. In conclusion, using statistical methods, the study informed the board on of its audience, the flaws of the website and proposed recommendations to address concerns.

Keywords: Bayesian networks; Google Analytics; machine learning methods; Markov chains; survival models; web analytics, web personas.

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Prof Retius Chifurira and Prof Temesgen Zewotir, for their valuable guidance in completing this research study, the many late nights spent on conference calls, and their assistance in putting together the five papers that we have worked on. Furthermore, I would like to thank the editor, Hester A. van der Walt.

I would also like to thank my wife, Christine, and sons, Micah and Elijah, for their love, support and patience during the years of intense study.

Revelation 10 vs 7: "But in the days of the voice of the seventh angel, when He shall begin to sound, the mystery of God should be finished, as He hath declared to His servants the prophets."

CONTENTS

Page

List of Figures **viii**

List of Tables **x**

Research outputs **xi**

Chapter 1: Introduction **1**

1.1 Introduction 1

1.2 Background 1

1.3 Web analytics literature review 3

1.4 Statement of the problem 8

1.5 Aim and objectives 9

1.6 Significance of the study 9

1.7 Contributions 10

1.8 Thesis outline 11

Chapter 2: The data and exploratory analysis **12**

2.1 Introduction 12

2.2 Data description 13

2.3 Descriptive statistics 17

2.4 Outlier analysis and missing data 22

2.4.1 Outlier analysis 22

2.4.2 Missing data 28

2.5 Concluding remarks 36

Chapter 3: Investigating a website traffic data-shift using artificial neural networks **37**

3.1 Introduction 37

3.2 Research methodology 40

3.2.1 Artificial neural networks 40

3.2.2 South African lock-down levels 46

3.3 Empirical results . . . 47
3.3.1 Online behaviour . . . 48
3.3.2 ANN feature selection . . . 49
3.3.3 Artificial neural net . . . 52
3.3.4 Feature importance . . . 54
3.4 Summary . . . 55

Chapter 4: Unsupervised machine learning feature selection on website data 57

4.1 Introduction . . . 57
4.2 Methodology . . . 59
4.2.1 Feature relevance . . . 59
4.2.2 Association between features . . . 62
4.3 Empirical results . . . 64
4.3.1 Measure of variability . . . 65
4.3.2 Measure of association . . . 66
4.4 Discussion . . . 72

Chapter 5: Identifying the underlying intentions of web visitors 74

5.1 Introduction . . . 74
5.2 Model description and methods . . . 75
5.2.1 K-means . . . 75
5.2.2 Hierarchical clustering . . . 77
5.2.3 Density-based spatial clustering of applications with noise . . 79
5.2.4 Cubic clustering criterion . . . 82
5.2.5 Silhouette coefficient . . . 83
5.2.6 Unsupervised modelling data preparation . . . 83
5.3 Empirical results . . . 84
5.3.1 Number of clusters . . . 84
5.3.2 K-means . . . 85
5.3.3 Hierarchical clustering . . . 88
5.3.4 DBScan . . . 93
5.4 Final thoughts . . . 96

Chapter 6: Survival models to identify the key hazards of website drop-offs 99

6.1 Introduction . . . 99

6.2 Materials and methods 100
6.2.1 Drop-off literature 101
6.2.2 Methodology flowchart 102
6.2.3 Kaplan-Meier curve 103
6.2.4 Cox proportional hazard regression 103
6.2.5 Random survival forest 106
6.2.6 Formal methods 109
6.3 Exploratory analysis 110
6.3.1 Webpage views 110
6.3.2 Kaplan-Meier curve 111
6.4 Survival models 112
6.4.1 Cox proportional hazard regression 112
6.4.2 Random survival forest 114
6.5 Final remarks 115

Chapter 7: Predict webpage transitioning using Markov chains 119

7.1 Introduction 119
7.2 Markov methodology 121
7.2.1 Discrete Markov chains 121
7.2.2 Transition matrix 121
7.2.3 Absorbing state 122
7.2.4 Variations of Markov chains 122
7.3 Markov states 123
7.4 Markov results 125
7.4.1 Webpage state historic dependence 125
7.4.2 Distribution of webpages viewed 126
7.4.3 Tiered Markov models 127
7.4.4 Joint (hidden) Markov model 130
7.5 Concluding thoughts 132

Chapter 8: Discussion and Conclusion 135

8.1 Introduction 135
8.2 Future directions 137
8.3 Limitations of the study 139

References **158**

LIST OF FIGURES

Figure 2.1 An illustration of the study's web analytics data pipeline. 12
Figure 2.2 TEKmation home page (source: Author's own contribution). 13
Figure 2.3 Web session metrics. 17
Figure 2.4 Web activity metrics. 19
Figure 2.5 Web device-specific metrics. 20
Figure 2.6 Web geographic metrics. 21
Figure 2.7 Total visits per client distribution. 23
Figure 2.8 Total visits per client distribution - post outlier removal. 24
Figure 2.9 Total visits per client distribution - histogram post outlier removal. . 25
Figure 2.10 Visit duration distribution. 26
Figure 2.11 Visit pageview distribution. 27
Figure 2.12 Directed Bayesian network model for imputation. 33

Figure 3.1 Artificial neural network architecture. 41
Figure 3.2 Architecture of an artificial neuron and a multilayered neural network 43
Figure 3.3 Feed-forward artificial neural network. 44
Figure 3.4 Fully recurrent artificial neural network. 46
Figure 3.5 Correlation matrix between features considered for the ANN. 49
Figure 3.6 Box plots of scaled features considered for the ANN. 51
Figure 3.7 Constructed artificial neural network. 53
Figure 3.8 Feature importance in identifying an online web data-shift. 54

Figure 4.1 Correlation matrix of numeric features prior to unsupervised machine learning. 67
Figure 4.2 a) user type x home, b) country x distance, c) user type x dayssincelastsession, d) user type x organic searches. 69

Figure 5.1 Sample dendrogram. 78

Figure 5.2 The cubic clustering criterion values. 84
Figure 5.3 Online web visit dendrogram. 89

Figure 6.1 Web analytics data flow chart. 102
Figure 6.2 Survival analysis pageview distribution. 110
Figure 6.3 Survival analysis Kaplan-Meier curve. 111
Figure 6.4 Web analytics survival Cox proportional hazards. 113
Figure 6.5 Web analytics random survival forest hazards. 114

Figure 7.1 Online visit page decision tree. 124
Figure 7.2 Studied website extract (source: Author's own contribution). 125
Figure 7.3 Distribution of the number of pages viewed per visit. 126
Figure 7.4 Tier one Markov chains. 128
Figure 7.5 Tier two Markov chains. 130

LIST OF TABLES

Table 1.1 Literature review summary. 7

Table 2.1 Web analytics original variables. 14
Table 2.2 Web analytics processed variables. 16
Table 2.3 Predicted vs actual operating system on the test dataset. 34
Table 2.4 Predicted vs actual mobile device brand on the test dataset. 35

Table 3.1 SA COVID-19 lock-down levels. 47
Table 3.2 Aggregate key online metrics at the various lock-down levels. 48

Table 4.1 Mean and measures of variability within numeric features. 65
Table 4.2 Measures of variability within categorical features. 66
Table 4.3 Box-and-whisker plot levels of associations. 70
Table 4.4 chi-squared test for independence: p-values. 71

Table 5.1 Notations used within the section. 75
Table 5.2 K-means cluster profiling across web metrics. 85
Table 5.3 K-means cluster profiling across webpage visits. 86
Table 5.4 K-means cluster summary. 87
Table 5.5 Hierarchical cluster profiling across web metrics. 90
Table 5.6 Hierarchical cluster profiling across webpage visits. 91
Table 5.7 Hierarchical cluster summary. 92
Table 5.8 DBScan cluster profiling across web metrics. 93
Table 5.9 DBScan cluster profiling across webpage visits. 94
Table 5.10 DBScan cluster summary. 95
Table 5.11 Web visit cluster comparison. 97

Table 6.1 Features identified as survival hazards. 116

Table 7.1 Tier one transition probabilities. 127
Table 7.2 Tier two transition probabilities. 129
Table 7.3 Joint Markov chain model of order two. 131
Table 7.4 General Markov chain model concerns. 133
Table 7.5 Average transition probability of the Markov models. 134

ABBREVIATIONS

ANN	Artificial neural network
API	Application programming interface
CCC	Cubic clustering criterion
COVID-19	Corona virus pandemic of year 2019
CPT	Conditional probability table
DBSCAN	Density-based spatial clustering of applications with noise
HTML	Hypertext markup language
IP	Internet protocol
RSF	Random survival forest
SA	South Africa
SMME	Small, micro, and medium enterprises

RESEARCH OUTPUTS

All journals considered for the publication were statistical journals indexed by Scopus. The following papers have been published or accepted for publication, stemming from this thesis:

1. **Soobramoney, J.**, Chifurira, R., Chinhamu, K., Zewotir, T. (2022). Selecting key features of online behaviour on South African informative websites prior to unsupervised machine learning. *Statistics, Optimization & Information Computing*, *11*, 519-530, doi: https://doi.org/10.19139/soic-2310-5070-1139

2. **Soobramoney, J.**, Chifurira, R., Zewotir, T. (2023). Modelling the South African COVID-19 induced web traffic data shift using artificial neural networks. *Applied Mathematics & Information Sciences*, *16*(6), 1049-1056, doi: https://dx.doi.org/10.18576/amis/160623

3. **Soobramoney, J.**, Chifurira, R., Chinhamu, K., Zewotir, T. (2023). Identifying the key hazards behind website drop-offs by solving a survival problem. *Applied Sciences*, *13*(14), 8248, doi: https://doi.org/10.3390/app13148248

4. **Soobramoney, J.**, Chifurira, R., Chinhamu, K., Zewotir, T. (2022). Identifying the intents behind website visits by employing unsupervised machine learning models. Accepted for publication by *Annals of Data Science*

5. **Soobramoney, J.**, Chifurira, R., Chinhamu, K., Zewotir, T. (2023). A tiered approach to Markov models when future events are not independent of the past: an application in web analytics. *Journal of Statistics Applications & Probability*, *S1* (Dec 2023), 1449-1458, https://dx.doi.org/10.12785/jsap/12S106.

CHAPTER 1

INTRODUCTION

1.1 Introduction

This chapter outlines the background of the study, related literature is reviewed, the problem statement, aim and objectives of the study are presented. Furthermore, this chapter details the significance, contribution of this study and concludes with the outlines of the subsequent chapters.

1.2 Background

The internet has transformed into a global marketplace (Hu and Haddud, 2020; Luo, 2021). Customers are able to purchase or browse goods and services that are globally available by using the world wide web, which can be accessed through devices such as mobile phones, tablets or computer devices (Mbete and Tanamal, 2020; Rita et al., 2019; Wang and Law, 2019). As a result, businesses are compelled to design and maintain website's that are attractive, efficient, user-friendly and competitive (Bleier et al., 2018; Noorbehbahani et al., 2019). In the context of online shopping, the corporate's website can be approximated to a store's branding and image (Miao et al., 2022; Tien, 2017). Although online shopping is the norm in first-world countries, developing countries are adopting this form of e-commerce rapidly (Haji, 2021; Khan et al., 2019; Mustafa et al., 2022; Tunio et al., 2023). Online shopping has proven to be mutually beneficial to shoppers and stores (Al-Lami, 2021; Xuhua et al., 2019). The latter stores benefit from reduced operating costs, reduced reach limitations, faster response to market demands, et cetera, whereas consumers benefit, among other things, from the convenience, no crowds and access to product reviews.

A website can be closely monitored by using web tracking tools/plug-ins such as Google Analytics. The tracking tools record the volume of users who

enter the website, their activity on the website, their online duration, the devices used to access the website, their geo-location, the number of times the person has entered previously, et cetera (Kantanantha and Awichanirost, 2022; Kumar and Ogunmola, 2019; Onder and Berbekova, 2021). Although these web data may be highly complex due to the variances in browsing behaviour, the data may also be highly insightful. Data scientists would seek to understand browsing patterns and ultimately seek to optimise spending and customer engagement (Mahmutovic, 2020; Sood, 2022). The study has employed traditional and modern statistical methods to make sense of the web tracking data of a corporate that has begun its digitalisation journey.

This study was conducted on a South African engineering and training corporate (TEKmation). TEKmation is classified as a small, micro, and medium enterprise (SMME) with several campuses across South Africa and footprints in Mozambique and Angola. The objectives of TEKmation are the following: Firstly to offer engineering consulting services to other entities or assist in the manufacturing or modification of products that require specialised engineering skills and tools. Secondly, to offer engineering training that can enable students to attain nationally recognised qualifications such as trade-test qualifications, apprenticeships, learnerships, and more. The trades offered by TEKmation include mechanical engineering, air-conditioning, refrigeration, electrical engineering, welding, instrument engineering and a series of other associated courses, for example basic-first-aid and health-and-safety.

The board of directors at TEKmation invested in a corporate website. The intention behind the website was to facilitate marketing of the corporate in the hope of growing awareness and, ultimately, revenue. At the time of the study, the website of TEKmation was an informative one, containing detailed information about the entity itself and its offerings (without a log-in, online sale or online study functionality). However, the board of directors was not informed of the traffic on the website, nor if it was used in the intended manner. Although the members of the board believed that the website visits could potentially be complex to make sense of at face value (due to endless possible page paths a visitor could follow), they were concerned about the volume of people entering the website, visitors' activity on the website and whether the drop-off rates were acceptable.

1.3 Web analytics literature review

Web analytics is often referred to as the study of a website's visitors, with a primary interest in comparing visit behaviour to the intended purpose of the website; in other words, a website is designed with a specific purpose, and website owners have to ensure that it is being used in the intended manner. Web analytics involves, by definition, tracking, reviewing and reporting on visits to understand web activity, including the use of a website and its features, for example, webpages, images and videos (Belair-Gagnon and Holton, 2018). By construct, a typical website is built off a domain (www.tekmation.co.za) that represents the main website; thereafter, several webpages are attached to the website (e.g., *about-us* or *contact-us*) and are often placed after the '\' domain (e.g., www.tekmation.co.za\about-us) (Huang et al., 2022; Schmitt et al., 2021). Although on a given domain, a visitor can navigate through the webpages that belong to that domain by clicking on available links.

The literature that is further discussed within Section 1.3 highlights related work on website analytics. Porsche et al. (2022) conducted a study to understand reading behaviour within online books. The main purpose of the study was to assess the performance of the Google Analytics tracking tool as a format suitable for advanced tracking of reading behaviour within web books, to prescribe measurements for reading the behaviour of web books, and to present the results of a pilot study. Using the analytics conducted, the study suggested deployment procedures of web books and presented possible methods of web book performance evaluation. Furthermore, the study concluded that the Google Analytics tool was a valuable tool for tracking traffic to individual books and quantifying the traffic to the entire web book collections on the observed data, by using unique custom and advanced metrics that were proposed.

Domazet and Simovic (2020) conducted a study to measure the performance of an online informal educational institution by means of web analytics. The goal of the study sought to determine the best-performing acquisition channel for non-formal educational institutions and the aggregate visitor profile of this kind of educational programmes by means of visitor acquisition and

behaviour data. The key metrics employed to assess the performance of the various acquisition channels were the conversion rate, average session duration, and bounce rate. However, visitor demographics (e.g., gender and age data) were supplemented on the side of the visitor's specific data. The findings of the study concluded favourable and suggested that the findings emerging from this study could apply to other non-formal educational institutions too.

Jonathan et al. (2019) employed web analytics to understand church members' activities on the Church-Cast application that hosted online sermons. The researchers believed that the study would ultimately increase user knowledge and interaction. Church-Cast was developed to make sermons of the Gospel ministries more available to their church members by means of digital channels of the internet and mobile devices. It was believed that time constraints and religious restrictions of low-capacity church members in certain parts of the world had resulted in church members being unable to attend their locations of worship physically to listen to or watch their ministers. The application, being web-based, tracked visitors' usage and thereby informed the administrator of visitors' activities on the application.

Semeradova and Weinlich (2020) examined the web traffic of user-friendly websites and proposed an analytical procedure based on data sourced from Google Analytics, using the online interface that is made available by Google. The researchers claim that web-user experience testing may often be very time-consuming, costly, and biased since the number of web-testers is frequently quite small. Within the study conducted by the researchers, a proposed analytical framework based on the web tracking data, sourced from Google Analytics, was presented. The framework proposed leveraging off the features and capabilities provided by the Google Analytics web-interface and introducing the concept of virtual pageviews as a user attention indicator. The study conducted A/B testing during which each test group represented an e-shop design. The performance of each e-shop design was analysed to determine the leading design.

Kalyankar and Anute (2022) sought to assess the value that web analytics yielded to e-commerce. The primary objective of the study was to identify how web analytics was employed within the e-commerce sector, as well as to

determine the way in which e-commerce enhanced a corporate's financials by means of web analytics. The study focused on target strategies and attempted to advocate web analytics in sales and marketing. The findings of the study highlighted the importance of first having strong financial goals set to make meaningful sense of web tracking data.

Cirlugea et al. (2020) performed web analytics to assess the impact and optimisation of Facebook advertisements on a small Romanian company that specialised in the manufacturing of clothing in an artisanal and artistic manner. The primary objective of the study was to gauge the performances of Facebook adverts in terms of customer reach, engagements and reactions to the adverts; to quantify the effectiveness of Facebook adverts towards the sale of the products; and to establish a possible target audience for the brand. The outcomes of the study were expected to illustrate a Facebook advert guide for young niche businesses to grow.

Rosqa and Ati (2022) conducted a study on an Indonesian corporate by using web analytics in conjunction with qualitative data obtained from the corporate to assess the vital statistics of the corporate's website (e.g., bounce rate and pageviews). The main objective of the study was to quantify the statistical results of the Elzatta website analysis by using a web tracking tool. The study employed a combination of a qualitative approach through observation data collected at the corporate's premises and secondary web analytics data sourced from the corporate's website. The findings of the study indicated that the corporate's visitors totalled 187,163 visitors within the six-month analysis period. Furthermore, the volume of website visitors increased monthly, with the most frequent visitor profile being females between the ages of 18 and 34 years from the Indonesian towns of Jakarta, Surabaya and Bandung. Most website visitors accessed the website through mobile devices that either entered the website through browsing for the website or following a social media link from Facebook and Instagram. Furthermore, the study recorded a high bounce rate of 51.48%.

Pirvu and Anghel (2019) proposed a double recurrent neural network machine-learning solution that reads visitors' behaviour in their previous visits and predicts their behaviour in their current visits. The researchers claim that e-commerce web applications are embedded within their own day-to-day lives

as well as the local economy. However, most applications are believed to lack the adaptation to users' needs and, subsequently, result in sub-optimal conversion rates and unsatisfied customers. The study presents a machine learning tool that learns the interests' of visitors from previous sessions and predicts useful metrics for the current session. Thereby, these predicted metrics can inform applications to allow customisation and allow better recommendations. This will also allow presenting better offers of specific products, targeted notifications or the placing of targeted adverts. Pirvu and Anghel believe that the proposed model will be accurate and allow applications to better customise the website offers to a client's needs and better predict a target base. The findings of the study indicated that the model can yield a probability score for each of the defined target classes.

Mariyapillai and Pratheepan (2021) conducted a web analytics study on an online library web portal of the Uva Wellassa University (UWU); their primary interest was to understand the spatial distribution of library visitors. The study tracked a few webpages, including the *home* page, online public access catalogue, and an institutional repository. The website analytics indicated that the studied website had been visited by roughly 366,756 local and global visitors during the period of the study. On their studied website, most visitors entered the University's website from the United States of America (15.82%), followed by visitors from the Netherlands (4.78%).

Stelian and Stoicu-Tivadar (2020) employed web analytics to study and quantify the level of interaction in a virtual reality (VR) medical application that was designed for educational purposes. The medical application allowed users to observe and handle human bones in a virtual environment. The web analytics indicated that the foot bones attained an interaction rate of 80%, whereas other skeletal structures attained an interaction rate of 100% on the observed data.

Table 1.1 provides a summary of the related work presented within this study.

Table 1.1: Literature review summary.

Authors	Context of study	Methods	Key findings/recommendations
Porsche et al. (2022)	Web analytics on online books.	Exploratory/no statistical methods.	Google analytics proved valuable in monitoring traffic to specific books and collections of books.
Domazet and Simovic (2020)	Web analytics on an online informal education institution.	Exploratory/no statistical methods.	The study was able to determine the best-performing acquisition channels.
Jonathan et al. (2019)	Web analytics on an online Church platform.	Exploratory/no statistical methods.	The study indicated that the church platform was useful for those unable to physically attend places of worship.
Semeradova and Weinlich (2020)	Web analytics study across several user-friendly websites using Google's Analytical dashboard.	Exploratory/no statistical methods.	Using Google's Analytical available reporting facilities, it was possible to conduct A/B testing to optimise web design.
Kalyankar and Anute (2022)	Web analytics of e-commerce.	Exploratory/no statistical methods.	The study highlighted the importance of web analytics within e-commerce.
Cirlugea et al. (2020)	Web analytics to assess the impact of Facebook adverts.	Exploratory/no statistical methods	The study illustrated a Facebook guide for small businesses.
Rosqa and Ati (2022)	Web analytics on an Indonesian e-commerce website	Exploratory/no statistical methods.	The exploratory analysis demographically profiled the client base.
Pirvu and Anghel (2019)	Web analytics recommender system.	Artificial neural networks to make recommendations based on learnings from the visitors' previous browsing patterns.	The ANN model proved to be highly accurate in recommendations to target groups.
Mariyapillai and Pratheepan (2021)	Web analytics on an online library.	Exploratory spatial analysis.	The study profiled the geo-demographics of the audience.
Stelian and Stoicu-Tivadar (2020)	Web analytics on a virtual reality application.	Exploratory/no statistical methods.	The findings of the study detailed the usage of the virtual reality components.

Given the recent literature presented in Table 1.1, none of the discussed literature has indicated analysis that guided the researcher on how the underlying intentions of a website were determined. Furthermore, several of the related literature employed descriptive statistics to aggregate the complex web visit

data. However, descriptive aggregation is less insightful relative to the methods detailed within this study.

Although it is evident that web analytics has attracted much research in recent years, there are still gaps in the literature that this study aims to fill. The research presented within this thesis differs from the literature in several ways:

- Firstly, the construct of the studied website is different. This study is conducted on a South African informative engineering website (i.e., a non-transactional website) in which the online behaviour will differ from most other websites, as the underlying intention behind a visit is dependent on the nature of the website under study.
- Secondly, at the time of writing, none of the reviewed literature applied statistical methods to explore, addressed concerns within the data, or extracted meaning from the complex web analytics data.
- Thirdly, this research addresses concerns that are specific to these data, for example investigation of the data-shift, as well as further investigation of the clusters and intents that emerged from the studied data. The high volume of drop-off intents was also studied by means of survival models to identify the key hazards and Markov chain models in order to study the transition probabilities of drop-offs, thus, filling a crucial gap in literature.

1.4 Statement of the problem

The directors of the TEKmation board have invested in their corporate website and, in return, sought to understand the following: firstly, what visitors were doing on their website; and secondly, whether there was any evidence of concerns that require immediate changes to the website. Although these seem to be practical questions, the answers thereof were certainly not straightforward. The nature of online browsing data is highly complex and the data are often large in volume (Awan et al., 2021; Fu et al., 2019). Visitors entering a website have differing intentions, follow unique page paths, spend varying amounts of time on it, may re-visit it days later, and so forth.

The study aims to understand visit behaviour by analysing the complex web browsing data and examining potential concerns to ultimately promote the growth of TEKmation within a competitive environment.

1.5 Aim and objectives

The main aim of the study is to employ statistical and machine learning methods on the complex web-visit data to yield a clear understanding of the underlying visit behaviour and identify any evidence of potential concerns on the studied website. This was achieved by

i quantifying the web-traffic volumes and assessing the data quality on missing data and outliers through exploratory analysis;

ii investigating the factors behind a data-shift using artificial neural networks;

iii feature selection prior to unsupervised machine learning across the various online web data types by using redundancy and relevance quantitative methods;

iv personifying and identifying the intents behind website visits by using clustering techniques;

v identifying the factors influencing the time to website drop-offs hazards by using random survival forest models and Cox proportional hazards models;

vi predicting the transition probabilities of a website visitor from the initial webpage to the next page, by using tiered Markov chain models.

1.6 Significance of the study

This study has been conducted on real-life data and the findings of the study aim to address the research objective of the organisation. The findings and recommendations of the study are to be embedded within the business strategy to grow its market share. Furthermore, given the rise in e-commerce (particularly in developing markets), this study can also be used as a guide

for corporates to understand their digital channels by using the statistical methods employed. A common problem faced by many website owners is elevated levels of website drop-offs (Garg and Dhari, 2019; Wang et al., 2021); the latter refer to the event when a person exits a website prematurely (Poulos et al., 2020). This study details methods to understand and address premature drop-offs. Due to the complex nature of website data, many similar applications, such as traffic theory, can also leverage off the analytical methods employed within this study. At the time of writing, limited research could be found documenting a detailed advanced analytical framework for website data.

1.7 Contributions

The major contribution of this thesis is the application of statistical and machine learning techniques in modelling online behaviour. The contributions are as follows:

1. The evaluation of statistical methods of feature selection prior to unsupervised machine learning models on web analytics data.
2. The assessment of the ability of artificial neural network models to detect shifts in browsing data patterns.
3. Studying the underlying intents behind web visits by using three fundamentally different unsupervised machine learning techniques.
4. The comparison of a Cox proportional hazard and a random survival forest model in their ability to detect the hazards that drive time to website drop-offs.
5. The exploration of tiered Markov chain models to predict the most likely subsequent webpage per visitor, given that the visitor was on a certain page.

The contributions have added value to TEKmation and have been published in five academic papers by statistical journals that are indexed by Scopus. At the time of writing, four papers have been published and one has been accepted for publication.

1.8 Thesis outline

The subsequent chapters are discussed as follows: Chapter 2 of this thesis introduces the information that the Google Analytics tracking tool records and presents the exploratory analysis on the observed data.

Chapter 3 employs artificial neural networks to understand the impact of the data-shift caused by the COVID-19 pandemic and has been published by the *Applied Mathematics and Information Sciences Journal*; the article is titled "Modelling the South African COVID-19 induced web traffic data shift using artificial neural networks".

Chapter 4 details feature selection methods for unsupervised machine learning on the website data types and has been published by the *Statistics, Optimization & Information Computing Journal*, it is titled "Selecting key features of online behaviour on South African informative websites prior to unsupervised machine learning".

Chapter 5 discusses the underlying intents that were identified behind the website visits on the observed data which discovered a concern with website drop-offs which has been accepted for publication by the *Annals of Data Science Journal* entitled "Identifying the intents behind website visits by employing unsupervised machine learning models".

Chapter 6 probes the high degree of website drop-offs by using survival models. This has been published by the *Journal of Applied Sciences* and is titled "Identifying the key hazards behind website drop-offs by solving a survival problem".

Chapter 7 employs tiered Markov chains to predict the next webpage viewed; it has been published by the *Journal of Statistics Applications & Probability* and is titled "A tiered approach to Markov models when future events are not independent of the past: an application in web analytics".

Finally, Chapter 8 concludes the study and discusses the limitations and future work.

CHAPTER 2

THE DATA AND EXPLORATORY ANALYSIS

2.1 Introduction

The data under study was collected through the Google Analytics tracking tool. To enable Google Analytics tracking, a Google Analytics account had to be created initially and a tag obtained. Thereafter, the tag needed to be embedded within the website's hypertext markup language (HTML) code so that the tracking tool could monitor visits and store the data linked to the corresponding Google Analytics account (Alhlou et al., 2016; Weber, 2015). When this has been done, a data pipeline was constructed (as depicted in Figure 2.1). A data pipeline represents the process that ingests raw data from various data sources and thereafter stores the data within the intended environment, for example a data lake or data warehouse (Helu et al., 2020; Mitchell et al., 2022).

Figure 2.1: An illustration of the study's web analytics data pipeline.

As illustrated in Figure 2.1, visitors entered the studied website by means

of a mobile, desktop, laptop, or tablet device. The Google Analytics tracking tool monitored the visitors' activities, stored them within a cloud server and linked the data to the associated Google Analytics account created. The researcher then accessed the tracking data through Google's application programming interface (API) by using the R data science programming tool.

By design, it was an informative website that served to inform visitors of the entity's purpose, offerings, history, contact information and clients. However, the studied website did not have a log-in feature, nor could visitors transact on the website at the time of study. Figure 2.2 depicts the website's *home* page at the time of study (prior to further planned modifications).

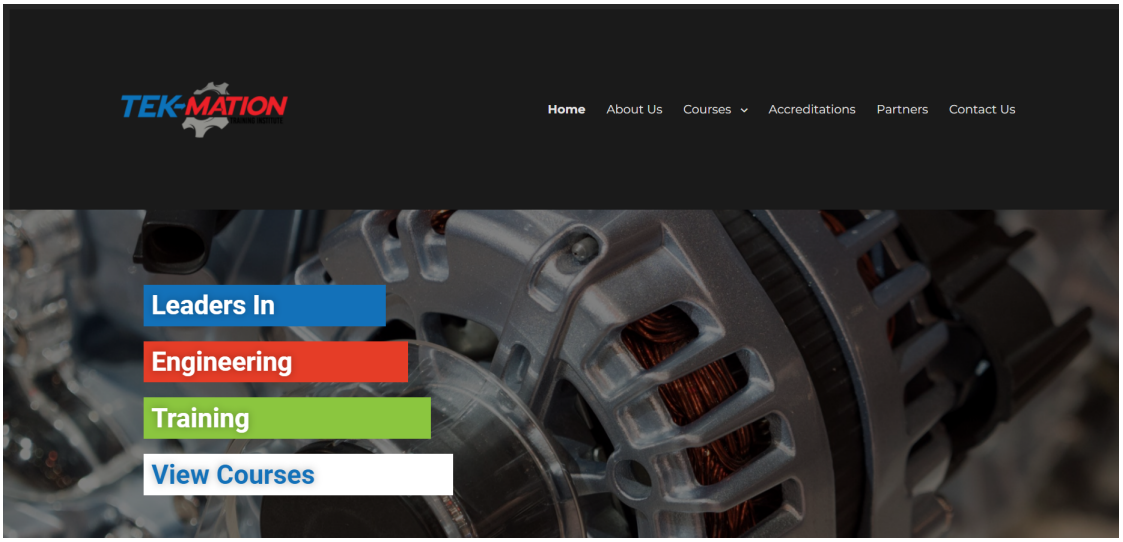

Figure 2.2: TEKmation home page (source: Author's own contribution).

The studied website contained 15 webpages in total, which could further be grouped into six high-level categories (as illustrated in Figure 2.2), namely *home*, *about-us*, *courses*, *accreditations*, *partners* and *contact-us*.

2.2 Data description

The data sourced from Google Analytics were recorded at a detailed level. Each visit was assigned a unique key and information about the visit was stored across several relational tables in a Microsoft SQL database as designed by the researcher (Meier and Kaufmann, 2019; Schule et al., 2021).

Table 2.1 discusses the variables that were available from the web tracking tool on the studied website.

Table 2.1: Web analytics original variables.

Web variable	Data type	Description
Bounces	Binary	Flag to indicate if a visitor has merely entered the website and thereafter immediately dropped-off.
Browser	Categorical	Indicates the web-browser used to visit the website (e.g., Chrome, Edge or Firefox)
ClientID	String	Represents Google Analytic's encrypted identifier (should a visitor return at a later point, Google Analytics would present the same identifier across all this person's visits). Therefore, although anonymous, the ClientID enables the study of repeat visits and the corresponding behaviour.
Country	Categorical	Indicates the country of the visitor's location when entering the studied website.
DateHourMinute	Datetime	Represents the time-stamp from the point of entry to the point of exit for each visit.
DaysSinceLastSession	Numeric	The number of days a user is returning to the website.
DeviceCategory	Categorical	Indicating if a tablet, mobile or desktop/laptop device was used.
Hits	Numeric	the count of any action taken on a webpage that results in data being sent to Google Analytics tracker (e.g., page-clicks).
Longitude/Latitude	Float	Indicates the geographic coordinates of the visitor when browsing the studied website.
MobileDeviceBranding	Categorical	Indicates the device brand (e.g., Samsung or iPhone).
OrganicSearches	Binary	Flag to indicate if a visitor entered the website by organically searching for keywords as opposed to clicking a link from a social media site or typing in the domain.
OperatingSystem	Categorical	Indicates the operating system of the device used to access the website (e.g., Android or iOS).
PagePath	String	Represents the webpage that the ClientID visited at a given point in time (e.g., *home* or *contact-us*).
Pageviews	Numeric	Represents the number of instances a page was loaded (or reloaded).
Region	Categorical	Indicates the geographical region that the visitor was in when entering the studied website.
Session	Numeric	Represents a visit and is made up of a collection of actions.
SessionDuration	Numeric	Represents the total duration of the visit (seconds).
Source	String	Represents the origin of traffic. Either a visitor has clicked on a social media link to enter the studied website or searched through a search engine to find the website (e.g., Facebook or Twitter).
UserType	Categorical	Indicates if the user is a new user or returning user.

Table 2.1 details the fields as recorded by the Google Analytics tracking tool on the observed website (Vecchione et al., 2016). The first column in the table lists the fields obtained; the data type column indicates the format of the corresponding fields, and the description column provides an explanation of the corresponding data field. These variables represent the raw data from the tracking tool. However, in order to achieve the study's objectives, some of these fields had to be further processed to extract more meaning from them;

for example, a distance was created by computing the Euclidean distance between the visitors' coordinates and the TEKmation's offices (Patel and Upadhyay, 2020; Wu et al., 2021). A detailed list of the processed/inferred variables that were created for the purposes of analytics and modelling can be seen in Table 2.2.

Table 2.2: Web analytics processed variables.

Measure	Data type	Description
AC	Binary	Flagged as 1 if the first page viewed was the 'Accreditations' page.
Accreditations	Numeric	Count of visits the user made to this page within each session.
AP	Binary	Flagged as 1 if the first page viewed was the 'Apprenticeships' page.
Apprenticeship	Numeric	Count of visits the user made to this page within each session.
AU	Binary	Flagged as 1 if the first page viewed was the 'About-us' page.
Bounce rate	Percentage	The proportion of total visits that leave the website seconds after entering.
CE	Binary	Flagged as 1 if the first page viewed was the 'Customised-Engineering' page.
CL	Binary	Flagged as 1 if the first page viewed was the 'Clients' page.
Contact-us	Numeric	Count of visits the user made to this page within each session.
Courses	Numeric	Count of visits the user made to this page within each session.
CR	Binary	Flagged as 1 if the first page viewed was the 'Courses' page.
CU	Binary	Flagged as 1 if the first page viewed was the 'Contact-Us' page.
Customised-engineering-trading	Numeric	Count of visits the user made to this page within each session.
DataShift	Binary	A flag to indicate the COVID-19 lock-down levels four and five when the data-shift was experienced.
Desktop	Binary	Flags if the website was visited using a desktop or laptop device (as opposed to a tablet or mobile device).
DesktopRate	Percentage	The rate of visits from a desktop device.
Distance	Numeric	The Euclidean distance between the user's coordinates and the company's coordinates (owner of the website).
EA	Binary	Flagged as 1 if the first page viewed was the 'Engineering-Academic' page.
Engineering-academic-studies	Numeric	Count of visits the user made to this page within each session.
Engineering-Trade	Numeric	Count of visits the user made to this page within each session.
ET	Binary	Flagged as 1 if the first page viewed was the 'Engineering-Trade' page.
H	Binary	Flagged as 1 if the first page viewed was the 'Home' page.
Home	Numeric	Count of visits the user made to this page within each session.
Huawei	Binary	Flags if the website was visited using a Huawei mobile device.
International	Binary	Flags if the user is South African or not.
L	Binary	Flagged as 1 if the first page viewed was the 'Learnership' page.
LocalRate	Percentage	The incidence of visits located within South Africa.
MobileRate	Percentage	The rate of visits from a mobile device.
NewUserRate	Percentage	On a given day, the incidence of new users whom there is no evidence of having visited the website before.
PublicHoliday	Binary	A flag to indicate if the visit occurred on a South African public holiday.
Samsung	Binary	Flags if the website was visited using a Samsung mobile device.
SC	Binary	Flagged as 1 if the first page viewed was the 'Short-Course' page.
sessionCount	Numeric	An indicator of the nth time the user has accessed the website.
Short-courses-skilled-programmes	Numeric	Count of visits the user made to this page within each session.
Survived	Binary	Flags the visits that have survived. By definition, observations were considered to have survived if they viewed three or more webpages within the visit to the studied website.
T	Binary	Flagged as 1 if the first page viewed was the 'Trade-test' page.
Trade-test-arpl	Numeric	Count of visits the user made to this page within each session.
U	Binary	Flagged as 1 if the first page viewed was the 'University-of-technology' page.
University-of-technology-uot	Numeric	Count of visits the user made to this page within each session.
Users per day	Numeric	The number of unique viewers that visit the website per day.
Weekday	Binary	A flag to indicate if the visit occurred on weekdays.
YearEnd	Binary	A flag to indicate if the visit occurred during the year-end period.

2.3 Descriptive statistics

This section presents the exploratory analysis to give a high-level overview of the study's web analytics data. The observed website data were stored in a Microsoft SQL database (Jameel et al., 2022; Shijitha et al., 2022) and a Power-Bi report was constructed to aggregate and visualise these data (Becker and Gould, 2019; Trieu et al., 2022). In the report that was developed, there were four main sections (as detailed further within this section), namely sessions, activity, devices and geography. Although web tracking was enabled since the start of 2019, the exploratory analysis below represents two years of web visits, from June 2021 to June 2023.

Figure 2.3 depicts the session-related visuals on the studied website and discusses the key session metrics.

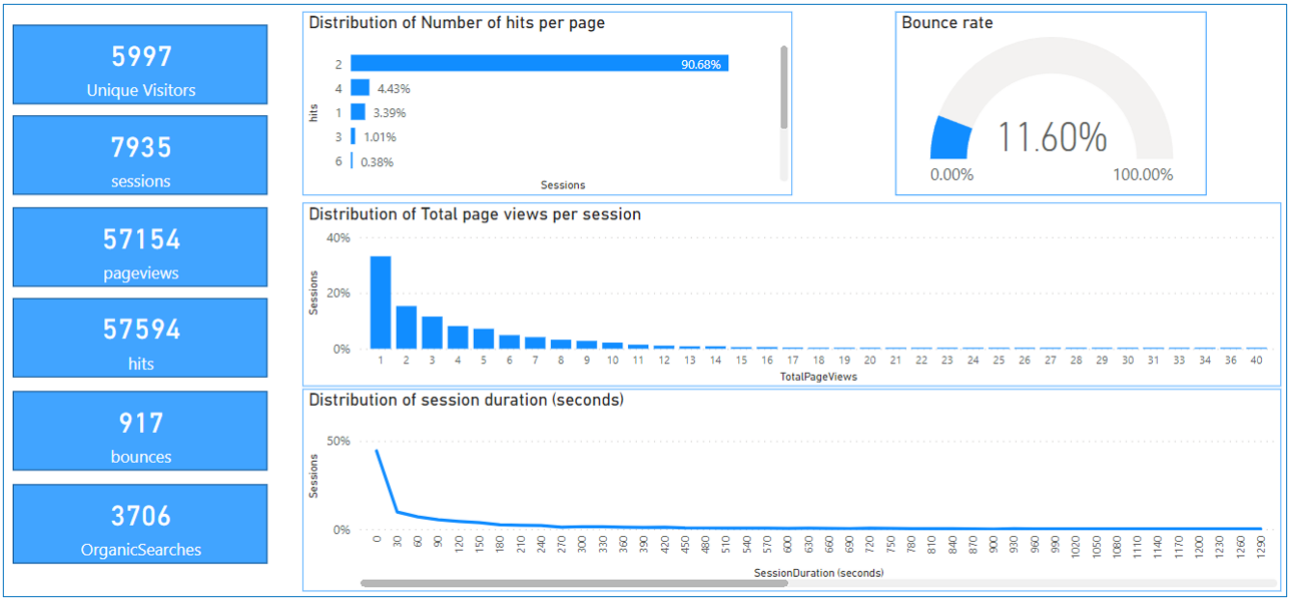

Figure 2.3: Web session metrics.

As depicted in Figure 2.3, the studied website received 5,997 unique visitors (as determined by the unique ClientID) who visited the website over the analysis period. The total number of visits on the observed data amounted to 7,935 sessions. This implies that the average unique visitor visited the website 1.32 times. The total number of pages viewed amounted to 57,154. The number of pageviews and hits (57,594) were almost equal, as by design, the studied website required mouse-clicks to navigate from one page to another.

The observed data recorded 917 bounced sessions. This represents visits during which a person would enter the website and exit immediately. Although the exact reason behind the bounce was unknown, it was likely due to a person entering by mistake (e.g., the visitor was looking for another company with a similar name). Therefore, the bounce rate on the observed data equated to 11.6%, which was computed by assessing the proportion of total sessions that had bounced (i.e., 917 bounces out of 7,935 visits). A high bounce rate could be indicative of serious website flaws. Jansen et al. (2022) conducted a study on 86 websites and observed that the average bounce rate across the studied websites equated to 56.2% with the lowest being 20.4% and the highest 88.9%. However, in the observed study, a bounce rate of 11.6% was acknowledged as satisfactory.

By assessing the distribution of hits per page (excluding bounced visits), it can be concluded that most people action at least two hits per page. Furthermore, roughly 33% of visits only view one webpage before exiting (inclusive of bounced visits and visits with a negligible amount of time spent online).

It was also observed that roughly 44% of the visits were less than 30 seconds in duration (excluding bounced visits). Although a satisfactory bounce rate was observed, it was concerning to realise that roughly one in three visits have only viewed one webpage prior to exiting the website (inclusive of bounced visits and visits with a negligible amount of time spent online).

Figure 2.4 discusses the repeat behaviour of visitors on the studied website. The tracking tool assigned a unique ClientID to a visitor and should this visitor return at any point in the future, the same ClientID would be used to enable studying the repeat behaviour of visitors.

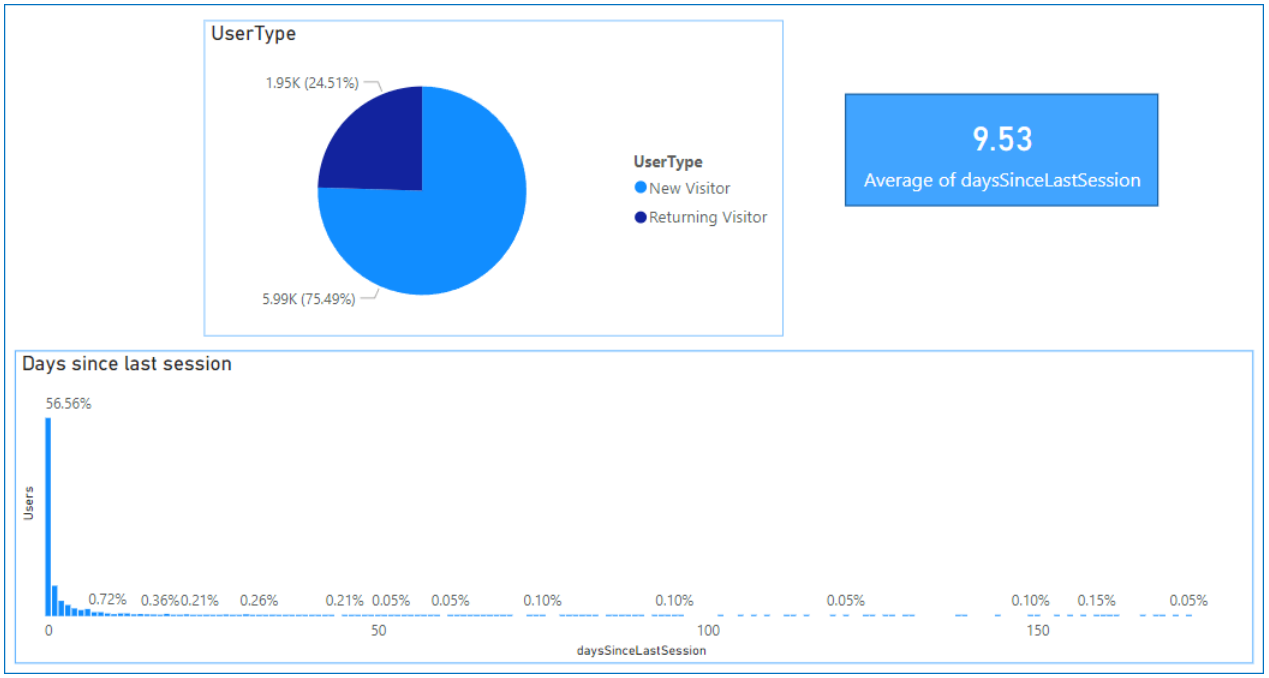

Figure 2.4: Web activity metrics.

As illustrated in Figure 2.4, the studied website was mostly visited by new visitors. Within the pie chart of the figure, *new-visitor* represents the volume of visits belonging to visitors who have accessed the website for the first time, whereas *returning-visitor* represents the volume of visits that belong to visitors who have accessed the studied website at least once before. Roughly, one in every four visits was from returning visitors who had accessed the website previously. Conversely, three in every four visits were from new visitors who had accessed the website for the first time.

Of the returning visitors who had accessed the website previously, the average return time on the studied website was 9.5 days between the previous visit and the current one. However, when assessing the detailed distribution in the number of days since the previous visits of the returning visitors, the observed study indicated that roughly 57% of returning visits occurred at a later point on the same day. From the right-skewed distribution (Atilgan et al., 2019) in the number of days since the previous visit, it could be indicative of the likelihood that returning diminished as time progressed.

The web tracking tool that was enabled on the studied website was able to record detailed information on the device-specific metrics (Figure 2.5). The

tracking tool was able to determine the type of device used (i.e., mobile device, desktop/laptop device or tablet device), the web browser used (e.g., Chrome or Safari), the operating system of the device (e.g., Windows or Android), and the device brand (e.g., Samsung or Huawei). Such level of information can be highly informative in cases where the website may contain compatibility flaws across specific devices. For example, if a particular website does not load correctly onto a specific web browser due to a compatibility error, the data will indicate no visits from such a browser or an outstanding drop-off rate (Bentameur and Belmihoub, 2022; McGuirk, 2023).

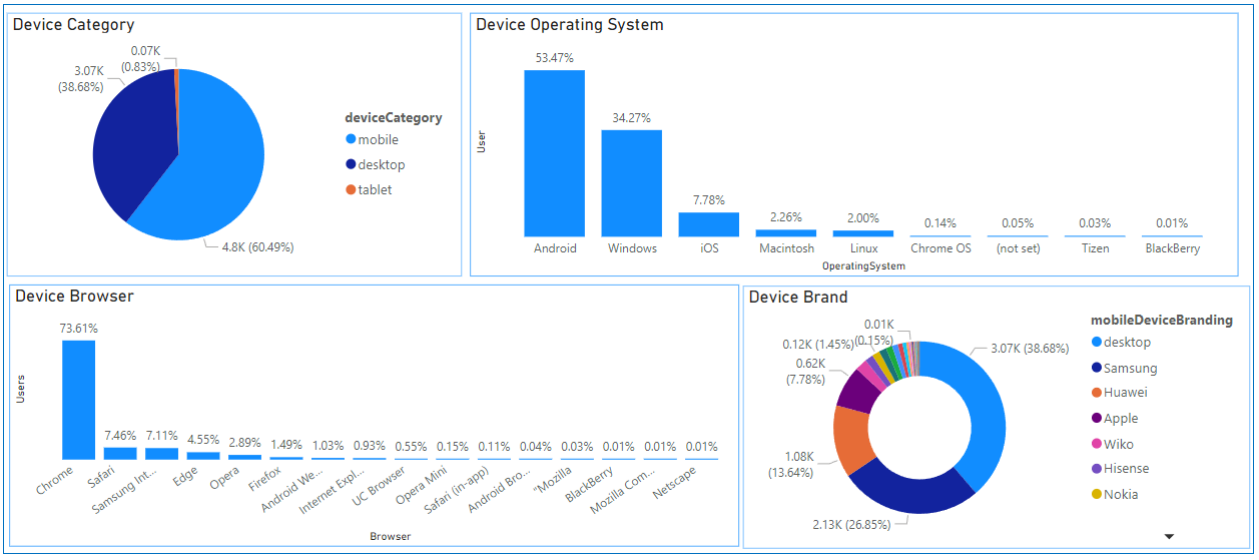

Figure 2.5: Web device-specific metrics.

As depicted in the pie chart in Figure 2.5, roughly 60.5% of visits to the studied website were through mobile devices. The portion of visits from desktop computers or laptops accounted for 38.7%, whereas tablet devices were used to access the website for just under 1% of the total visits. Google Chrome was observed as the most frequent web browser (73.6%) used by visitors to enter the studied website, followed by Safari (7.5%) and Samsung Internet (7.1%). Although the operating system is dependent on the type of device used (Crutcher et al., 2021), the studied website observed android systems being the most used operating system (53.5%), followed by Windows (34.3%), iOS (7.8%), and Macintosh (2.3%). Furthermore, Samsung has featured as the most popular mobile device (26.9% of all visits) used to access the studied website, followed by Huawei (13.6%) and Apple (7.8%).

The tracking tool enabled on the studied website was able to determine the geographic latitude and longitude coordinates from the visitor's IP (internet protocol) address. The latter represents a unique numerical identifier, assigned by the internet service provider, for every device or network connected to the internet (Waggoner et al., 2019).

Therefore, as illustrated in Figure 2.6, the spatial distribution of visits could be assessed at a country and regional level to understand the geographic location of visitors on the studied website. Some devices (due to anti-virus or personalised web browser settings) may prevent the tracking tool from determining the detailed geographical location of the visitor, though.

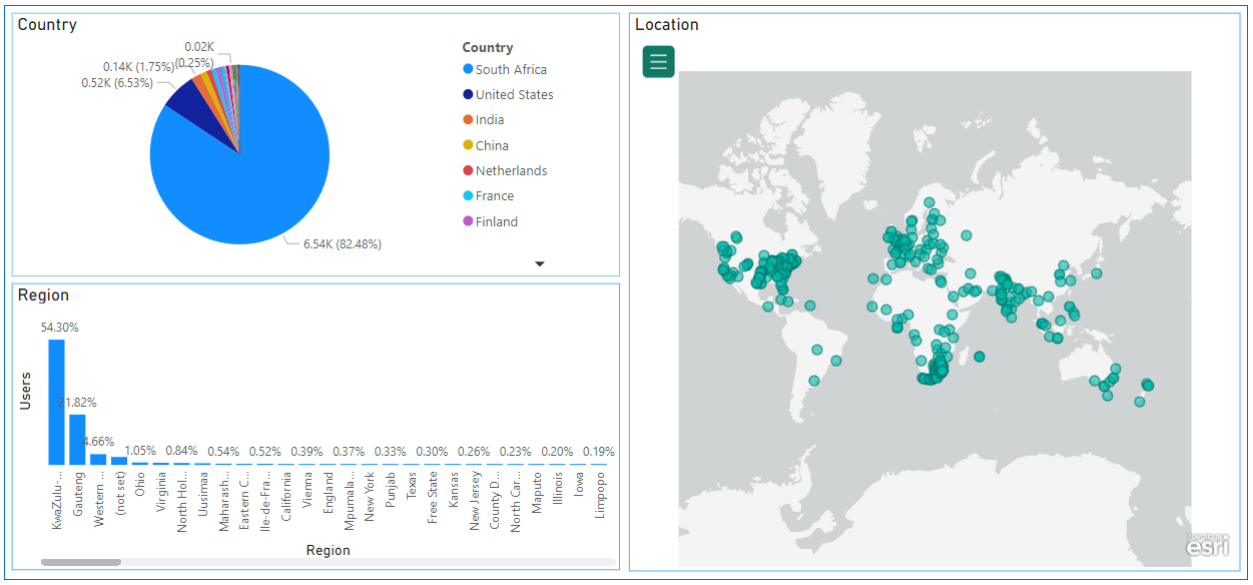

Figure 2.6: Web geographic metrics.

According to the studied website, (as illustrated in Figure 2.6), most visits were from South Africa (the country the studied corporate belongs to). However, a material portion (17.5%) of visits belonged to foreign countries, with the United States accounting for 6.5% of total visits and India accounting for 1.8%. At a regional level, the most frequent region was KwaZulu-Natal (54.3%), the province where TEKmation's head office was based and where it held its largest footprint. Subsequently, the trailing regions within South Africa were Gauteng (21.8%) and the Western Cape (4.7%). It can also be seen that the portion of visits from unknown regions was minimal (as indicated by the *not-set* label) and represented 3.5% of the total visits. However, the

country information could still be determined for the region with *not-set* visits. The spatial distribution in Figure 2.6 depicts the cities from which there were visits and provides a visual of the concentration areas of interest.

2.4 Outlier analysis and missing data

This section discusses the assessments conducted to identify extreme value data points and quantify the levels of missingness on the studied data. Furthermore, the section discusses the actions followed in response to the outlier and missing data assessments conducted on the data.

2.4.1 Outlier analysis

Outlier analysis can be described as the process of identifying and probing the data points that significantly differ from the general trend of the data (Ali Mohammadi and Chen, 2021; Buschjager et al., 2022). Outliers often represent extreme values found within quantitative data types that can distort statistical measures such as the mean values (Shrifan et al., 2021). However, two schools of thought prevail; although researcher agrees that outliers are harmful to aggregated metrics, many believe that outliers represent true and real-world occurrences and should therefore not be ignored (Andre, 2021; Uzun Ozsahin et al., 2022).

On the observed web tracking data, the presence of outliers was investigated on the total numbers of website visits per ClientID, the total number of pages viewed per visit and the total visit duration per visit. Grubb's test for outliers indicated a p-value of 0.00, which implied the existence of outliers (at a 5% level of significance) on the total number of website visits per ClientID (Uba et al., 2021).

Figure 2.7 depicts a scatter plot of the total number of visits per ClientID.

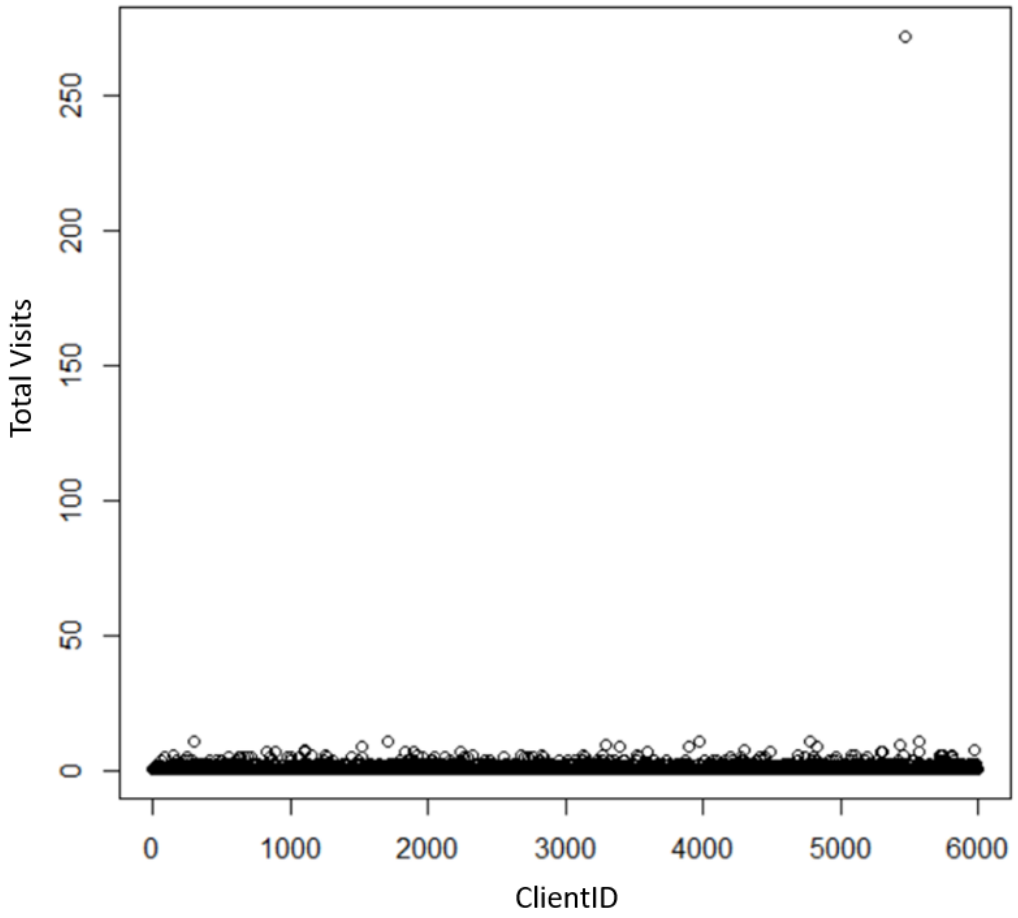

Figure 2.7: Total visits per client distribution.

The x-axis in Figure 2.7 represents the individual visitors. whereas the y-axis represents the corresponding number of total visits to the observed website (Santos et al., 2021). It was apparent that an extreme value existed. The tracking tool recorded that a specific ClientID had visited the website 272 times. Given that the analysis period spanned over two years, further investigations indicated that this visitor had visited the website daily and could possibly represent a communal desktop or a staff member. Upon studying the behaviour of this outlier during each of the individual visits, the browsing pattern did not raise any alarms.

However, since this extreme value suppressed the remaining data points, it was removed in Figure 2.8 to allow the researcher to study the remaining distribution of visits per ClientID.

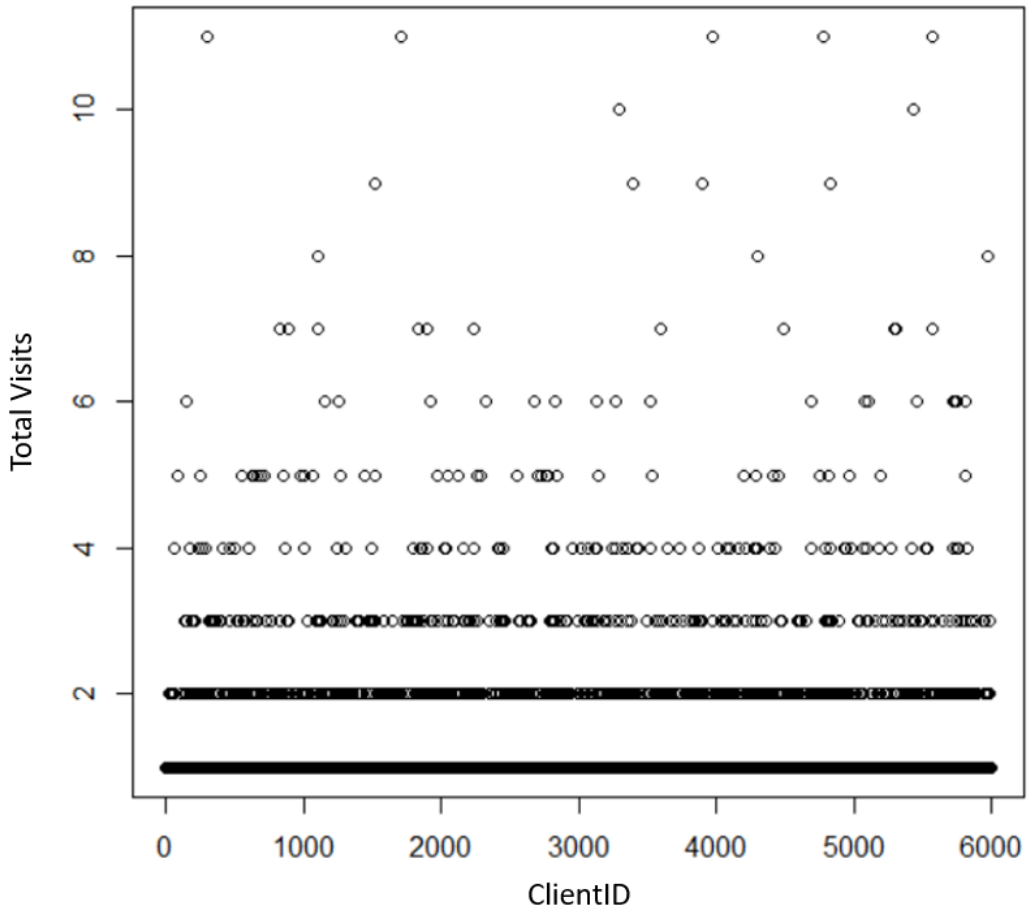

Figure 2.8: Total visits per client distribution - post outlier removal.

Figure 2.8 represents the scatter plot of the number of visits to the website per ClientID after the removal of the extreme value that was identified in Figure 2.7. The x-axis of Figure 2.8 represents the individual visitors and the y-axis the corresponding count of visits. It was then observed that the total number of visits per person ranged between one and eleven times. It inference was also drawn that the lower the visit frequency, the higher the concentration of data points.

Figure 2.9 thus represented the histogram that quantified the volume of unique visitors per frequency. This indicated the volume of ClientID's that visited the website once, twice, three times and more within the analysis period of the studied website.

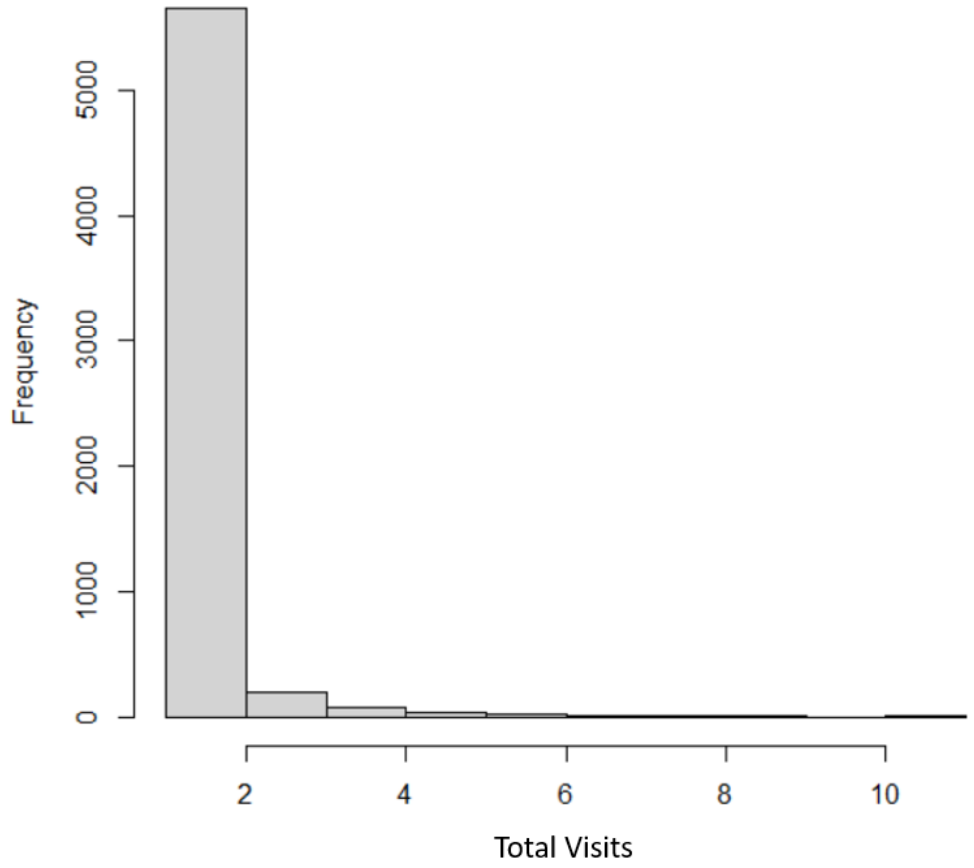

Figure 2.9: Total visits per client distribution - histogram post outlier removal.

The histogram depicted in Figure 2.9 indicates that the vast majority of unique ClientIDs visited the studied website once within the analysis period. The x-axis represented the count of total website visits, whereas the y-axis represented the count of unique ClientIDs (Nino Rondon et al., 2022; Wang et al., 2022). Thereafter, the volume of unique ClientIDs that visited the studied website on a more frequent basis diminished steeply. The histogram presented in this figure reflects the distribution after the removal of the identified outlier to allow for a better understanding of the remaining data points.

Subsequently, the distribution of the time each visitor spent on the studied website per visit was assessed to identify the existence of extreme values (as depicted in Figure 2.10). Grubb's test for outliers indicated a p-value of 0.01, which implied the existence of outliers (at a 5% level of significance) on the visit duration per visit (Uba et al., 2021).

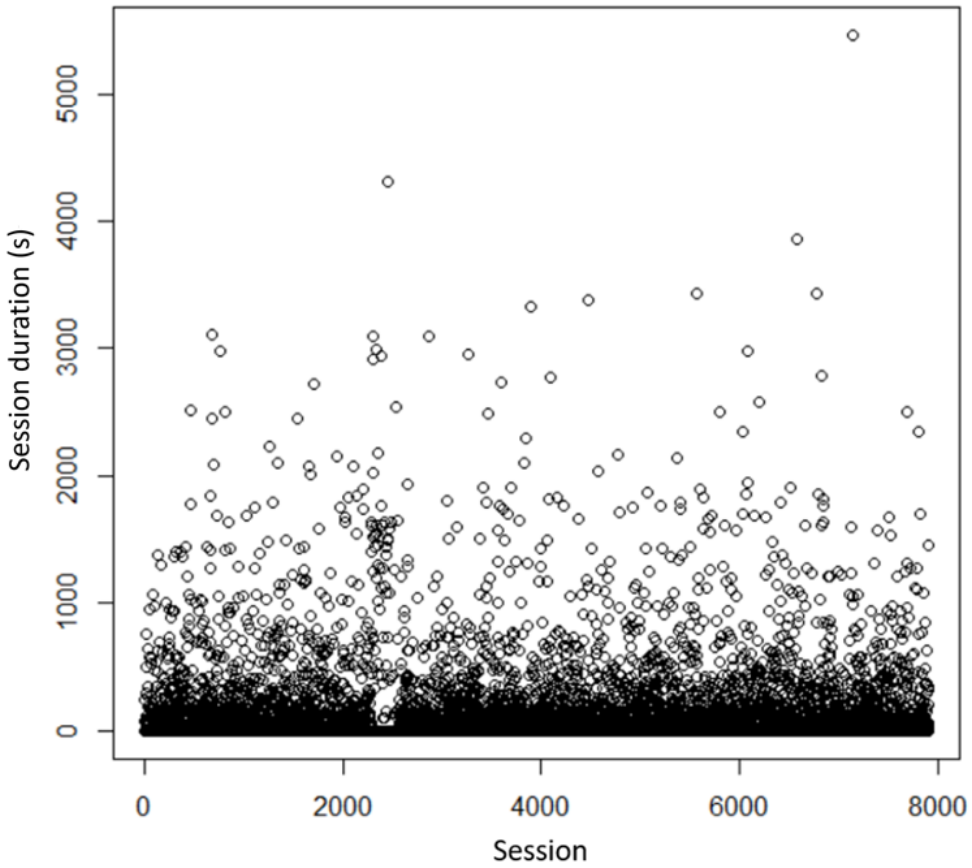

Figure 2.10: Visit duration distribution.

Figure 2.10 represents the total visit duration per visit in seconds, where the x-axis of the scatter plot denotes the individual visiting records, and the y-axis represents the subsequent time spent on the studied website during that visit. It was observed that although most of the visits were less than 600 seconds (10 minutes), there were diminishing volumes of observations that reached up to 5,460 seconds (91 minutes). Although the visits with a duration of over 4,000 seconds differed materially from the rest of the observations, the implications were still within reason, as the researcher believed it was understandable that a visitor could spend up to 90 minutes on the studied website if they wanted to read all the provided content properly.

The third measurement involved the collection of data on the number of pages viewed per visit; these data were then assessed for extreme values. The web tracking tool that was enabled on the studied website tracked the non-unique count of pages viewed per visit. For instance, if, within a single visit, the visitor would view the *home* page, navigate to the *courses* page and

then return to the *home* page, the *home* page would count as two pages, since it was viewed twice. Grubb's test for outliers indicated a p-value of 0.00, which implied the existence of outliers (at a 5% level of significance) on the count of webpages viewed per visit (Uba et al., 2021). Figure 2.11 depicts the scatter plot of the total pageviews on the studied website.

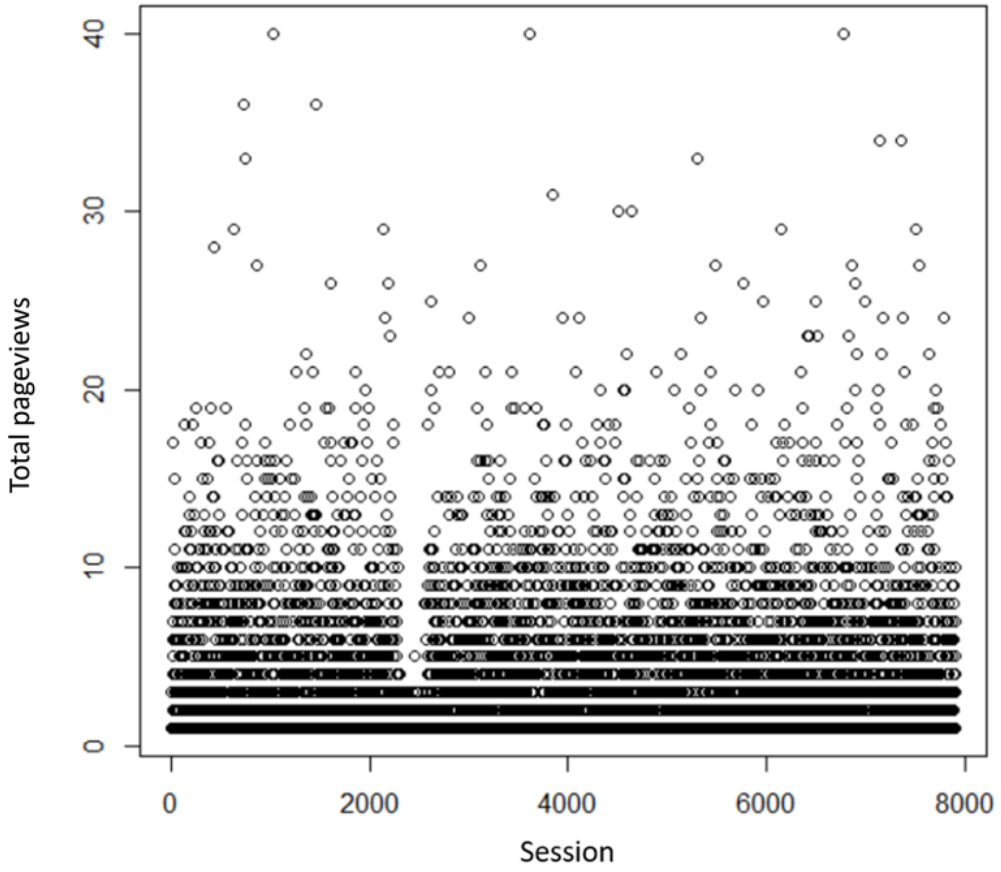

Figure 2.11: Visit pageview distribution.

The x-axis of Figure 2.11 represents the individual visits, and the y-axis indicates the corresponding count of non-unique pages viewed on the studied website. According to the distribution observed, most of the visitors viewed less than 10 pages, although there were some observations that reached 40 non-unique pageviews. However, the board of directors believed that it was reasonable for visits to reach up to 40 pageviews, as the website contained 15 unique pages.

Given the three quantitative aspects screened for outliers (visits per ClientID, duration per visit and pageviews per visit), all outlying data points were

found to be understandable and explainable on the studied website. However, should any of these measures be used in aggregation (e.g., K-means model), a winsorisation clean would be necessary (as discussed in Section 5.2.6).

2.4.2 Missing data

The engineering and training corporate, TEKmation, employed a web tracking tool to study the traffic flow and behaviour of users on the corporate website. However, in doing so, the web tracking tool yielded a small portion of incomplete information due to technical incompatibilities or resistance from certain applications that restricts sharing of certain information (e.g., geographic or device-related information).

Types of missingness

Missing data are believed to contribute to a loss of information and therefore result in a reduced statistical power (Enders, 2023; Gomer and Yuan, 2022). Furthermore, missing data have been found to potentially introduce selection bias and therefore invalidate data in extreme cases (Hong and Lu, 2023; Zhou et al., 2023).

The occurrence of missing values can be classified as either one of three types:

- Missing completely at random (MCAR) occurs when the missingness is entirely independent of all other features (Kularia et al., 2023).
- Missing at random (MAR) occurs when the missingness depends on the observed features. For example, respondents may not answer a certain question, since because they are uncomfortable with sharing their residential address information (Takai and Hayashi, 2023).
- Missing not at random (MNAR) refers to situations where the missingness of a feature depends on hidden causes (unobserved variables) (Pereira et al., 2023).

Within the study, the missingness was best described as missing completely at random, since the missingness had nothing to do with the person being studied and occurred mostly due to technology-related constraints. Furthermore, since the data had been tracked and compiled systematically, a low level of missingness was to be expected. Accordingly, the missingness could be attributed to technical incompatibilities or resistance from certain applications that hindered tracking of certain information.

Observed levels of missingness

Of all features recorded by the web tracking tool on the studied website, the mobile device brand and operating system features specifically contained missing information on the observed data. The degree of missingness on the mobile device brand was 1.24% and the operating system was 0.03%. Upon inspecting the individual visits that contained such missing data points, it could be seen that all other information was recorded. Therefore, instead of deleting these data points, the researcher attempted to impute the missing information (Chattopadhyay et al., 2023; Chowdhry, 2023). Although the percentage of incompleteness was minimal, this section discusses the use of Bayesian network models to complete the missing information.

Imputation using a Bayesian network model

Bayesian networks, also commonly known as probabilistic networks, Bayes nets or belief networks, belong to the family of probabilistic graphical models (Bibartiu et al., 2023; Vomlel et al., 2023). The term 'Bayesian networks' arose in 1985 and were further explored in the late 1980s by Judea Pearl and Richard E. Neapolitan. Pearl (1988) made use of the term 'Bayesian networks': *"A probabilistic network (also referred to as a belief network, Bayesian network or, somewhat imprecisely, causal network) consists of a graphical structure, encoding a domain's variables and the qualitative relationships between them, and a quantitative part, encoding probabilities over the variables"* (Ben-Gal, 2008).

Bayesian network definition

A Bayesian network is a graphical representation of a probability distribution over a set of variables in the universe

$$\Omega = X_1, X_2, ..., X_n.$$

It is composed of two facets (Cintron, 2022; Yao, 2023):

1. a directed network structure (S) in the form of a directed acyclic graph (DAG), and
2. a set of the local probability distributions (P) for each node (variable), conditioned on all possible combinations of the parental nodes.

Network structure

The network structure is constrained to be acyclic. This simply means that the structure does not loop or lead back to itself. Upon this graph, cycles may be undirected. An undirected cycle would have edges pointing in different directions. Such occurrences would represent possible influence between certain variables in the cycle.

Let Pa_i denote the parents of the corresponding variables at the nodes X_i and S_i. Then, the joint probability distribution function (pdf) for X is:

$$P(x) = \prod_{i=1}^{n} P(x_i|pa_i). \tag{2.1}$$

For any set of random variables belonging to Ω, the probability of any member of this joint pdf can be calculated from the conditional probabilities using the chain rule:

$$P(X_i = X_1, ..., X_n = x_n) = \prod_{\omega=1}^{N} P(X_\omega = x_\omega|X_{\omega+1} = x_{\omega+1}, ..., X_n = x_n). \tag{2.2}$$

As a result, the joint pdf can be written as:

$$P(X_i = X_1, ..., X_n = x_n) = \prod_{\omega=1}^{N} P(X_\omega = x_\omega|X_j = x_j), \tag{2.3}$$

for each X_j which is a parent of X_ω.
These probabilities may originate from two sources: It can be obtained from either prior knowledge or from the data itself.

Conditional probability tables

The conditional probability tables (CPTs) represent the quantitative component of the Bayesian network nodes. The CPTs define the conditional probability distribution of each node within the Bayesian network.

The derivation of CPTs can reach intensive levels due to the potential size of conditional probability distributions. Suppose a variable $T_i \in \mathbf{T}$, with a number of states r_i within T_i and the number of states r_j of each parent $T_j \in Pa(T_i)$ computed as:

$$size(CPT)_i = r_i . \prod_{r(T_j)} r_j . \tag{2.4}$$

Suppose r_j is equal for all T_i with n parent nodes, then the expression (Equation 2.4) reduces to:

$$size(CPT)_i = r_i . {r_j}^n . \tag{2.5}$$

The intensity arises since the size of the CPTs are exponentially related to the number of parents, n (as expressed in Equations 2.4 and 2.5). Hence, Bayesian models may be computer-intensive when models have numerous nodes.

Inference

Inferring with Bayesian networks involve utilising the CPTs to compute the probability of one or more nodes **T** given some prior knowledge or evidence, **e**. This is expressed as $p(T|e)$. Using Bayes' rule, $p(T|e)$ may be evaluated as follows:

$$P(t|e) = \frac{P(e|t).P(t)}{P(e)}. \tag{2.6}$$

Equation 2.6 is further expanded using the sum rule and product rule:

$$\frac{P(e|t).P(t)}{P(e)} = \frac{\sum_I (P(e|ti).P(i)).P(t)}{\sum_T \sum_I P(e|ti).P(t).P(i)}. \tag{2.7}$$

The denominator of Equation 2.7 serves as a normalisation constant that ensures the probabilities of T values sum to 1. Hence, a normalisation term is accounted for:

$$\alpha(P(e|t).P(t)) = \alpha(\sum_{I}(P(e|ti).P(i)).P(t)). \tag{2.8}$$

Although the computational math may be intricate, modern computers enable efficient and automated inference (e.g., the *bn-learn* R package).

This study employed Bayesian network models to impute the missing information identified within the observed web tracking data. The initial assumption was that the features with missing information were dependent on all other available features. Figure 2.12 depicts the assumed dependencies of the given study.

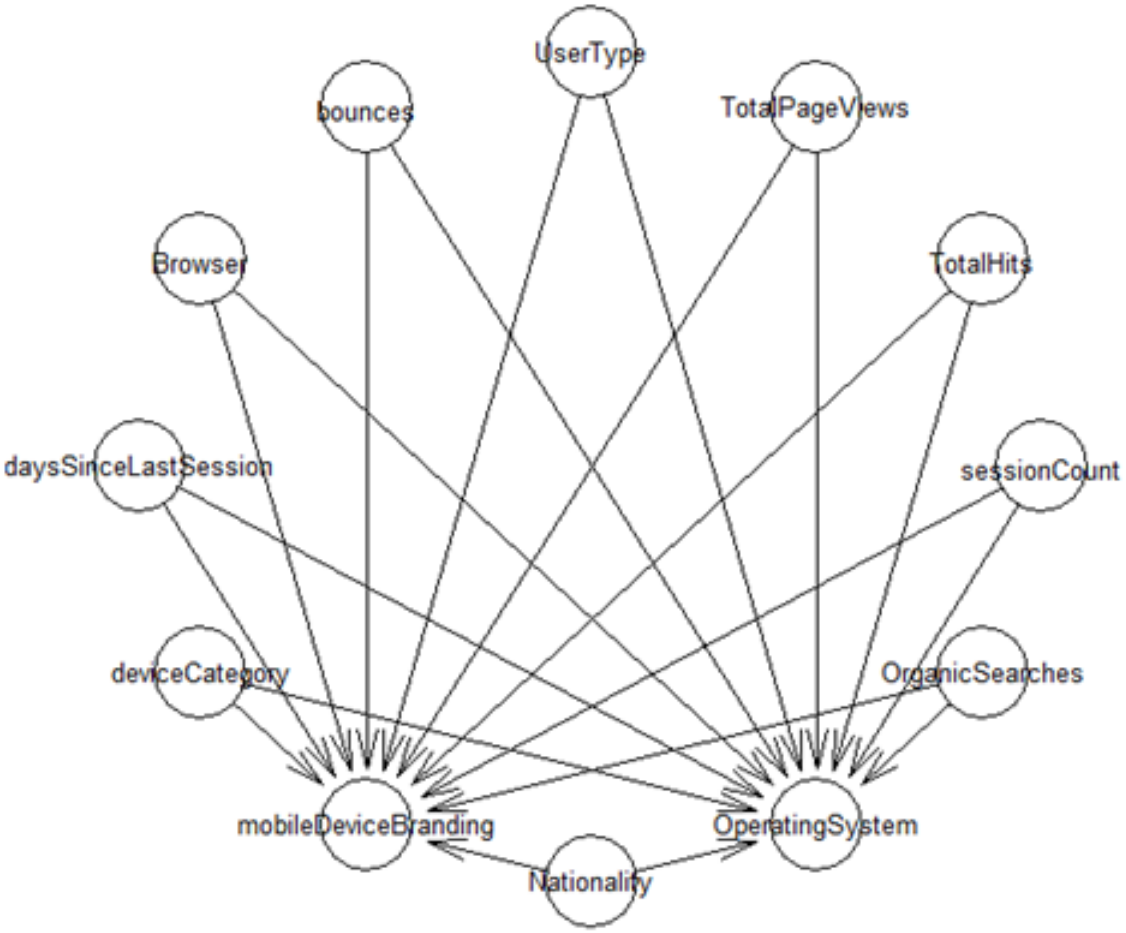

Figure 2.12: Directed Bayesian network model for imputation.

Within the observed data, the mobile device branding and operating system features recorded missing information. A Bayesian network structure was assumed in which the mobile device branding and operating system were dependent on all other categorical features available within the observed data. The dependency assumption would allow the model to construct the conditional probabilities that estimate the likely values of the mobile device branding and operating system, based on the subsequent values of all other known categorical features.

To assess the accuracy of this assumption, a subset of complete data was used to fit the assumed structure and estimate the conditional probabilities (the training dataset), whereas a subset of data was used to evaluate the accuracy of the fitted model by comparing the actual values to the predicted

values that resulted from the conditional probabilities (test data). On completed records, a random 95% of the data were used to train the Bayesian network model (Collart and Guisan, 2023; Singh et al., 2021), whereas a random 5% of completed records were taken to test the accuracy of the assumed dependencies and the conditional probabilities (Picardi, 2022). Due to the vast number of categories, the modelling allowed for a larger training base during the model build.

To gauge the accuracy of the Bayesian network's ability to impute the missing values, the predicted values were compared to the actual values on the validation dataset. Table 2.3 presents the confusion matrix on the operating system feature (Markoulidakis et al., 2021; Xu et al., 2019).

Table 2.3: Predicted vs actual operating system on the test dataset.

OperatingSystem	Android	iOS	Macintosh	Windows
Android	360	0	0	0
Chrome OS	0	0	0	0
iOS	1	29	0	0
Linux	0	0	0	4
Macintosh	0	0	11	5
Tizen	0	0	0	0
Windows	0	0	0	349
Windows Phone	0	0	0	0

In Table 2.3, the rows represent the actual data, whereas the columns represent the predicted values on the test data. A classification accuracy of 98% was attained on the test data (Mohammed et al., 2019; Moreno-Barea et al., 2020). Upon further investigation, the conditional probabilities detected strong associations between the operating system and the browser type (e.g., Chrome and Safari) or the type of device. For example, the iOS operating systems used were mainly the Safari web browser and mainly on Apple devices. Upon inspecting the low-incidence data points (operating systems with less than 30 visits), the confusion matrix illustrated that the Bayesian network models had performed poorly relative to the high-incidence data points.

Table 2.4 presents the confusion matrix of the mobile device brand feature to assess the accuracy of the predicted values from the Bayesian network model compared to the actual values on the test data.

Table 2.4: Predicted vs actual mobile device brand on the test dataset.

MobileDeviceBranding	Apple	desktop	Hisense	Huawei	Mobicel	Nokia	Samsung	Sony
Acer	0	0	0	0	0	0	1	0
Apple	29	0	0	0	0	0	0	0
Blackview	0	0	0	0	1	0	0	0
desktop	0	369	0	0	0	0	0	0
Hisense	0	0	2	9	0	0	0	0
Huawei	0	0	0	102	0	0	25	0
LG	0	0	0	4	0	0	1	0
Mobicel	0	0	0	0	1	0	0	0
Nokia	0	0	0	5	0	3	5	0
Samsung	0	0	0	62	0	0	142	1
Vivo	0	0	0	1	0	1	0	0
Vodafone	0	0	0	0	0	0	1	0

In Table 2.4, the rows represent the actual data, and the columns the predicted data. The classification accuracy on the mobile device brand feature was 85% on the test dataset. Here, the Bayesian network model had also detected an association between the device brand, the operating system and the browser. As illustrated in this table, the imputation results of low-incidence brands (mobile device brands with less than 30 visits) were relatively poor in comparison with the high-incidence brands.

To address the missing data detected on the studied website, a Bayesian network model was employed to impute the missing web tracking data (Santos et al., 2022; Vazifehdan et al., 2019; Ye et al., 2019). With web traffic tracking data, the volume of missingness was minimal and likely driven by technical incompatibilities or restrictive applications that prevented the tracking tools from fetching certain information. The Bayesian network model had a high overall classification accuracy on the test data when imputing missing mobile device branding and operating system features. The accurate classification of the mobile device branding and operating system features was attributed to strong associations between these features and the corresponding browser type and device category features. However, across both the mobile device

branding and operating system, the imputation results were often incorrect on low-incidence data points relative to the high-incidence data points.

2.5 Concluding remarks

Chapter 2 has discussed the features supplied by the Google Analytics web tracking tool on the studied website. A description of the features and data types were discussed. The exploratory analysis was also conducted to quantify the traffic on the studied website during the analysis period. This analysis indicated that the website was visited by 5,997 unique visitors, with 7,935 total visits and 57,154 pageviews. The studied website also indicated a relatively low bounce rate (11.6%) and of the 7,935 visits, 3,706 visits were found organically (46.7%). It was also established that most visits belonged to new visitors and a fair mix of mobile and computer devices were used to access the studied website. Although the website attained a wide geographic spread of visitors, the majority belonged to the major South African provinces. Although the observed bounce rate of the study was satisfactory, a major concern emerged about the material portion of visitors that merely viewed less than three pages on the studied website and thereafter dropped-off.

The data quality assessments included outlier analysis, using Grubb's test, on the number of visits per unique client, the visit duration per visit, and the number of pages viewed per visit. Although some observations were materially higher than the majority, all extreme values were rational according to the business owner's understanding of the studied website. Furthermore, a small degree of missing values was observed on the studied data within the mobile device branding and operating system features. With regard to the observed data, a Bayesian network model displayed satisfactory results on its ability to impute the missing data points.

Outside the analysis period of the study (June 2021 to June 2023), a material data-shift was observed prior to 2021 and is further discussed in Chapter 3.

CHAPTER 3

INVESTIGATING A WEBSITE TRAFFIC DATA-SHIFT USING ARTIFICIAL NEURAL NETWORKS

3.1 Introduction

Although the exploratory analysis presented in Chapter 2 spanned June 2021 to June 2023, the web tracking data was enabled from October 2019. However, going into the year 2020, a material data-shift was observed on the studied website. The exploratory and subsequent analysis was conducted on the data periods after the recovery of data patterns. Chapter 3 investigates the data-shift and proposes an artificial neural network model that could be used to continuously monitor the data patterns and alert if a shift in data patterns are detected.

To contain the spread of the coronavirus contagion of the year 2019 (COVID-19), the South African government attempted to balance economic and individual activity. Upon imposing harsh individual restrictions, the result would imply higher levels of unemployment and, conversely, too loose individual restrictions would result in a higher spread of the contagion than the healthcare systems could support (Kabisa et al., 2021; Kalogiannidis, 2020). The South African government managed the balance through a risk-adjusted strategy by introducing lock-down levels ranging from lock-down level one to lock-down level five. Lock-down level five allowed minimal economic activity and stronger isolation rules whilst lock-down level one permitted most industries to trade with the weakest individual isolation rules (Arora et al., 2020; Olivier et al., 2020). During the peak COVID-19 periods, global

economic health has been severely impacted across many sectors of industry (Baycan and Tuysuz, 2022; Mudahemuka et al., 2023; Selvanathan et al., 2021).

This chapter sought to investigate the shift in online user-behaviour during the peaks of the COVID-19 pandemic. The objective of the study was to identify the key factors of change observed on an informative website of an engineering training and engineering service provider (Bhrammanachote and Sawangdee, 2021; Galhotra and Dewan, 2020; Kumar et al., 2022). Furthermore, the chapter proposed an artificial neural network model that can be employed to detect the presence of future data-shifts based on the learnings of the COVID-19 induced data-shift. In the event of another data-shift occurring in the future, without being detected, the data-driven business decisions could be misleading. Thereby, it is suggested, that the proposed model be run in parallel with other data-driven tools such as reports and machine learning models that guide business decisions to detect a data-shift as soon as it occurs (Xiong et al., 2020).

Online behaviour change has been of interest to many researchers. Rodda et al. (2018) studied online behaviour change to understand problematic gambling. The researchers concluded that using change strategies to influence online problematic gambling behaviour was very complex and required further research with a broader population base (Rodda et al., 2018).

Perski et al. (2016) studied user engagement with digital behaviour change interventions. The researchers proposed a conceptual framework where the digital behaviour change intervention was influenced by the intervention itself. The researchers claimed that the context and mechanisms used may moderate the influence on the user's response to the digital behaviour change intervention (Perski et al., 2016).

Richiello et al. (2021) conducted a study to understand the challenges influencing webchat counselling. The researchers identified several possible methods to address the challenges and initiate a behaviour change. The study made use of an online behaviour change wheel and embedded surveys to address challenges (e.g., to prompt online counselling). However, it was found that low survey take-up rates may imply an impractical approach (Richiello et al., 2021).

The literature discussed further shares research on shifts in data patterns (data-shifts) across several other applications apart from online web data. Furthermore, some researchers highlight the danger that data-shifts impose on existing machine learning and analytical frameworks.

Kiley and Vaisey (2020) studied population-wide cultural change. The first aspect studied stated that people are actively updating their beliefs and behaviours as they process new information. The second aspect argued that following early socialisation experiences, the dispositions are stable. The data used were sourced from the General Social Survey. The study revealed that whilst short-term changes were noticed, persistent change occurred primarily amongst young people (Kiley and Vaisey, 2020).

Stacke et al. (2020) have claimed that neural network machine learning models are very accurate within a stable data environment. However, in the presence of a data-shift, unseen data poses a challenge to the generalisation of neural networks. The study focused on data-shifts within image processing employed within histopathology. The study proposed the use of convolutional neural networks to identify data-shifts thereby suggesting potential in-accuracies (Stacke et al., 2020).

Taori et al. (2020) assessed the impact of natural distribution shifts in image data compared to current synthetic distribution variations on predictive model robustness. The findings of the study concluded that distribution shifts that arise from real-world data remain an open research problem. This implied that such data-shifts do impact machine learning accuracy (Taori et al., 2020).

In an application of machine learning, to identify new biomarkers, Dockes et al. (2021) have highlighted the impact of data-shifts. The study conducted defined the breaking point and manner in which machine learning extracted biomarkers fail in the presence of data-shifts (Dockes et al., 2021).

Adams-Cohen (2020) used Twitter data and machine learning methods to study the shift in sentiment when the Supreme Court legalised same-sex marriages. The data-shift indicated that the Supreme Court's decisions polarise public opinion in the short term (Adams-Cohen, 2020).

Guo et al. (2020) proposed a method of enhancing neural network model accuracy in the presence of a data-shift. The study developed a continuous kernel cut segmentation algorithm by factoring in normalised cuts and continuous regularisation over the data. The outcome of the study proved that the method reduced segmentation variability and achieved excellent classification accuracy (Guo et al., 2020).

Xiong et al. (2020) emphasised on the hyper-parameter settings of machine learning models. In particular, the study proposed a protocol for assessing the hyper-parameter sensitivity to data-shifts. However, the results of the study indicated that no clear winner could be determined (Xiong et al., 2020).

This chapter assesses the key features of the COVID-19 induced data-shift and further proposes an artificial neural network (ANN) model that can be employed to detect the occurrence of a future data-shift based on learnings from the COVID-19 induced data-shift. At the time of the writing, no such literature existed in this context. Although online behaviour during the pandemic would be specific to the nature of the website itself, the approaches employed within this study can nonetheless be adopted on other websites to understand the significance of behavioural changes during the COVID-19 pandemic.

3.2 Research methodology

This study employed artificial neural network models to understand the data-shift in online web behaviour experienced during the COVID-19 outbreak. The models were used to understand the key factors of change detected within the data. Furthermore, a predictive model has been proposed to detect the occurrence of a data-shift should one occur in the future. Although other predictive models could have been likewise employed (e.g., ensemble methods), the artificial neural network has proven adequately accurate within this application.

3.2.1 Artificial neural networks

An artificial neural network (ANN) is a statistical machine learning algorithm that imitates the process of the human brain. An ANN consists of

artificial neurons that exist in several layers. These neurons are linked to each other through a network of connections (or nodes) (Hopfield, 1988). The ANN structure processes the data from inputs to an output using biases allocated to the neurons, weights associated with the connectors and an activation function. Figure 3.1 illustrates a three-layer artificial neural network. As illustrated in Figure 3.1, the input layer neurons are connected to the hidden layer neurons via weights. These weights dictate the influence that each input layer neuron imposes onto the hidden layer (Hopfield, 1988).

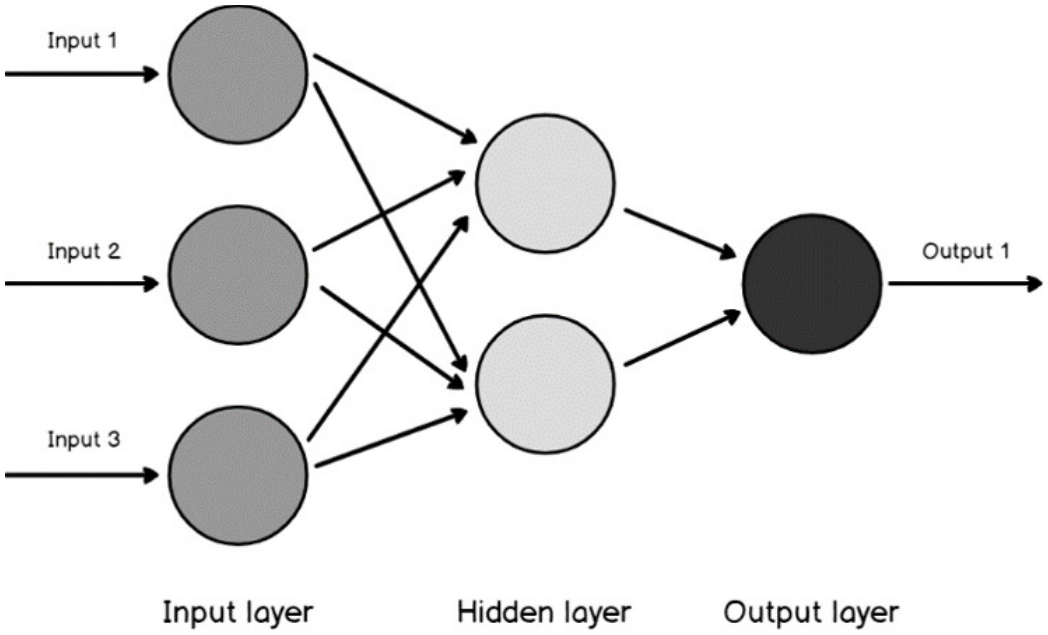

Figure 3.1: Artificial neural network architecture.

An ANN is optimised through the number of hidden layers and nodes that the modeller must determine on a given set of data. The ANN is able to determine the nodes that are more important than others by assigning more influential weights to the more important nodes (Hopfield, 1988). Although artificial neural networks are commonly employed in supervised problems and known to be high-performance models, ANN may also be implemented in unsupervised applications (Belhaj et al., 2021). There are several computational algorithms available to construct an ANN (e.g., multi-layer feedforward, feedforward or feedback networks). However, multi-layer feedforward models are considered the most frequently used architecture (Belhaj et al., 2021).

Following a general machine learning approach, modelling ANN consists of the collection of inputs and outputs (independent and dependent features), the selection of the model design (feed-forward or feed-backward), training the model and lastly, testing and validating the model. Depending on the outcome of the model, the process may be re-run by optimising the hyper-parameters (number of hidden nodes and hidden layers) for optimal model accuracy (Hopfield, 1988). Within the learning process of the model, the algorithm initially has a large prediction error when comparing the actual results to the predicted results. However, the system then minimises the error by adjusting the nodal weights through an iterative process. This iterative process of optimising the nodal weights is completed through back-propagation between the output nodes and hidden layers of the network. During the training process of the artificial neural network, a validation step is applied to optimise the performance of the model by governing the number of epochs per training cycle. The epochs represent the iteration process used by the model to minimise prediction errors of the model (Hopfield, 1988).

The final step of the modelling entails testing the model's performance on unseen data. By comparing the model's prediction to the actual (or desired) results accuracy metrics are quantified (Hopfield, 1988).

ANN definition

By definition, an artificial neural network (ANN) represents a generalisation of mathematical models of the biological nervous systems (Abraham, 2005). A typical artificial neuron (a) and the modelling of a multi-layered (b) neural network is illustrated in Figure 3.2.

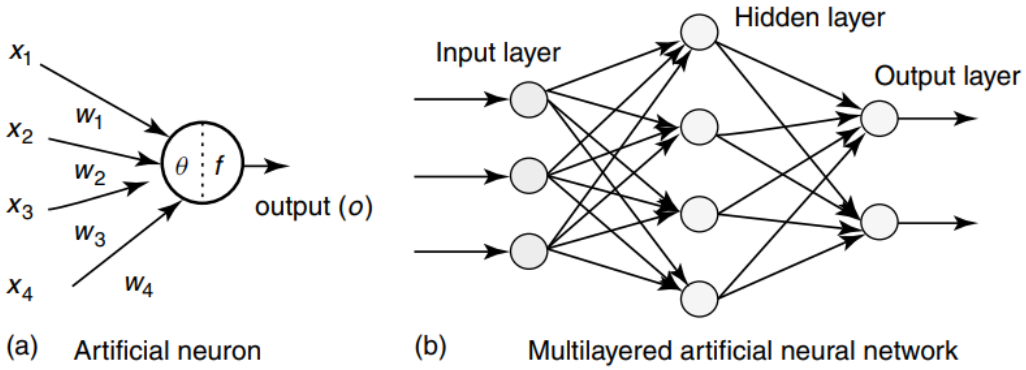

Figure 3.2: Architecture of an artificial neuron and a multilayered neural network.

With reference to Figure 3.2, the signal flow from inputs x_1,...,x_n is considered to be unidirectional, which are indicated by arrows, as is a neuron's output signal flow (O). The neuron output signal O is given by the following relationship (Abraham, 2005) as per Equation 3.1:

$$O = f(net) = f\left(\sum_{j=1}^{n} w_j x_j\right), \tag{3.1}$$

where w_j represents the weight vector, and the function $f(net)$ represents an activation (transfer) function. The variable net is defined as the scalar product of the weight and input vectors as per Equation 3.2:

$$net = w^T x = w_1 x_1 + ... + w_n x_n, \tag{3.2}$$

where T represents the transpose of a matrix and, in the simplest case, the output value O is computed as per Equation 3.3:

$$O = f(net) = \begin{cases} 1 & if \quad w^T \geq \theta \\ 0 & if \quad otherwise \end{cases}, \tag{3.3}$$

where θ represents the threshold level, with this type of node being termed a linear threshold unit.

Feed-forward artificial neural networks

By construct, a feed-forward artificial neural network represents an ANN with a forward feeding topology (Figure 3.3) and maintains the condition

that information must flow from input to output solely in one direction without back-loops (Krenker et al., 2011). The feed-forwarding construct holds no limitations on the number of layers, the type of transfer function employed within individual neurons or the number of connections between the individual artificial neurons. In the simplest form, a feed-forward artificial neural network contains a single perceptron that is known to learn solely linear separable problems.

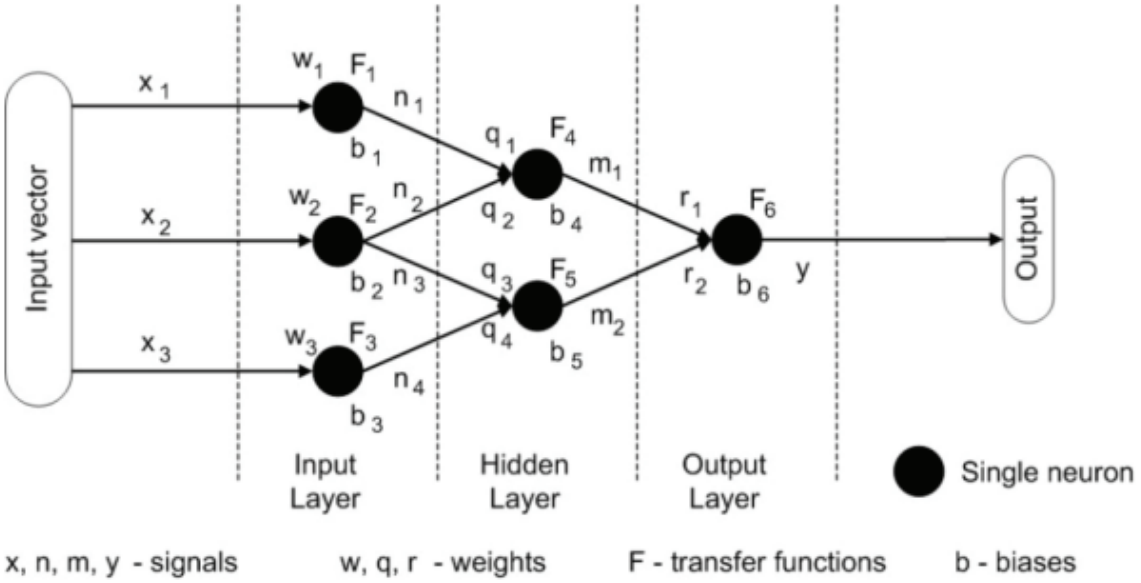

Figure 3.3: Feed-forward artificial neural network.

Where x,n,m and y represent signals; w, q, r represent weights; F represents transfer functions and b represents biases so that:

$$\begin{aligned} n_1 &= F_1(w_1 x_1 + b_1), \\ n_2 &= F_2(w_2 x_2 + b_2), \\ n_3 &= F_2(w_2 x_2 + b_2), \\ n_4 &= F_3(w_3 x_3 + b_3). \end{aligned} \tag{3.4}$$

$$\begin{aligned} m_1 &= F_4(q_1 n_1 + q_2 n_2 + b_4), \\ m_2 &= F_5(q_3 n_3 + q_4 n_4 + b_5). \end{aligned} \tag{3.5}$$

$$y = F_6(r_1 m_1 + r_2 m_2 + b_6). \tag{3.6}$$

$$y = F_6 \begin{bmatrix} r_1(F_4[q_1 F_1[w_1 x_1 + b_1] + q_2 F_2[w_2 x_2 + b_2]] + b_4) + ... \\ ... + r_2(F_5[q_3 F_2[w_2 x_2 + b_2] + q_4 F_3[w_3 x_3 + b_3]] + b_5) + b_6 \end{bmatrix}. \tag{3.7}$$

As illustrated in Figure 3.3, and as quantified by Equations 3.4, 3.5, 3.6 and 3.7 the simple illustrated feed-forward artificial neural network implies potentially long mathematical expressions. Therefore, in practice, specialised software is employed to build, explain and optimise simple and complex artificial neural network structures (Krenker et al., 2011).

Recurrent artificial neural networks

A recurrent artificial neural network, by construct, follows a recurrent topology as illustrated in Figure 3.4 (Krenker et al., 2011). A recurrent ANN, in principle, is similar to a feed-forward neural network however holds no limitations on back-loops within the flow process. With a recurrent ANN, information is no longer uni-direction but it is also transmitted backwards. Therefore, an internal state of the network is created which allows for a dynamic temporal behaviour. Figure 3.4 hypothetically illustrates a fully recurrent ANN and illustrates the complexity of the artificial neurons' interconnections. Within this illustration, every neuron is directly connected to every other neuron in all directions. There are several other recurrent artificial neural network constructs such as Hopfield, Elman, Jordan, bi-directional and other networks that are merely variations of recurrent artificial neural networks (Krenker et al., 2011).

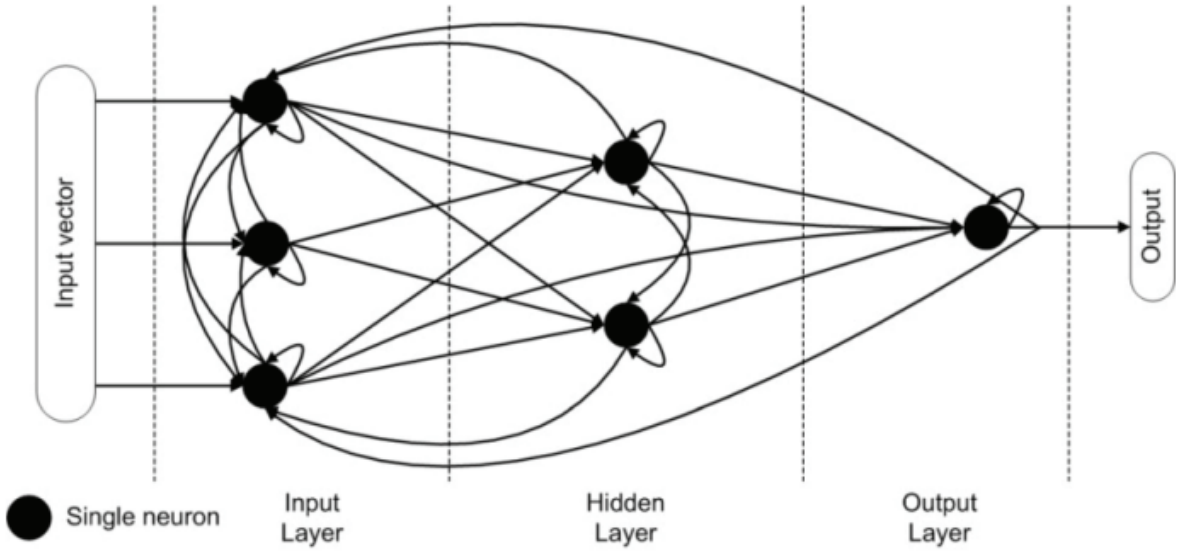

Figure 3.4: Fully recurrent artificial neural network.

3.2.2 South African lock-down levels

The online metrics supplied by the Google Analytics platform were observed across the South African lock-down levels that influenced the observed data-shift. The South African authorities regulated population and industrial behaviour through the implementation of lock-down levels at appropriate stages of the pandemic. The lock-down levels, as detailed in Table 3.1, sought to govern the trade-off between economic activity and the spread of the contagion.

Table 3.1: SA COVID-19 lock-down levels.

SA lock-down level	Description
Level 0 (L_0)	No economic or personal restriction (the period prior to the COVID-19 outbreak).
Level 1 (L_1)	Most normal activity can resume, with precautions and health guidelines followed at all times. The population prepared for an increase in alert levels if necessary.
Level 2 (L_2)	Physical distancing and restrictions on leisure and social activities to prevent a resurgence of the virus.
Level 3 (L_3)	High degree of precaution and restrictions on personal activities.
Level 4 (L_4)	Extreme precautions to limit community transmission and outbreaks, while allowing some activity to resume.
Level 5 (L_5)	Drastic measures to contain the spread of the virus and save lives.

The South African COVID-19 lock-down level five represented the most strict economic and socio-economic restrictions on the country. Level five was activated during the early stages of the outbreak whilst emergency hospital facilities were in development and when the spread of the virus was outrageous. And accordingly, between the waves, when the number of active COVID-19 cases was minimal, lock-down level one was activated at the lowest. It was found that during the harsh levels (four and five) of lock-down, a distinct data-shift was observed in the studied data. Of the observed features, several measures were directly sourced from the Google Analytics tracking tool, and other features were implicitly created (for instance, the *YearEnd* feature, which flags behaviour during the December holiday period to differentiate between seasonality and a data-shift). The data-shift feature flags the periods of lock-down level four and five when the data-shift was experienced on the observed data (Tables 2.1 and 2.2).

3.3 Empirical results

This section discusses the exploratory analysis of the studied online web behaviour to illustrate the data-shift. Furthermore, the outputs of an ANN model and subsequent feature importance are presented.

3.3.1 Online behaviour

This section discusses key aggregate online metrics between the various lock-down levels to assess the shifts in online behaviour. The analysis focuses on, firstly, traffic to the website and thereafter, the behaviour whilst on the website. Table 3.2 presents the average number of users visiting the website per day, the average session duration and the average number of pageviews per session (by definition, a session refers to a user's visit to a particular website).

Table 3.2: Aggregate key online metrics at the various lock-down levels.

	Avg. users per day	**Avg. session duration (s)**	**Avg. pageviews per session**
0. Level 0	5.39	387.80	3.32
1. Level 1	4.64	342.95	3.26
2. Level 2	4.97	331.76	3.52
3. Level 3	5.18	287.61*	3.21
4. Level 4	3.57*	298.23*	3.02
5. Level 5	2.83*	358.61*	3.10

From Table 3.2 for instance, on the observed website, an average of 5.4 users visited the website per day prior to lock-down and spent on average, 388 seconds during a typical session. During the harshest lock-down level five, an average of 2.8 users visited the website per day and spent, on average, 359 seconds on the website.

Using the Mann-Whitney U two-tailed test for significance (*), lock-down levels 4 and 5 record a significant difference in user-visits per day relative to level 0 (at a 5% level of significance). The test for significance further informs that the session duration during levels three, four and five are significantly different relative to the pre-COVID period. Although the session duration shows a statistically significant difference, the extent of the change seems minimal. It is evident that during the harsh levels of lock-down (levels four and five), the studied website experienced a drastic drop in traffic flow into the website. However, whilst on the website, user-behaviour remained fairly

stable across the lock-down (as quantified by the pageviews and session duration).

3.3.2 ANN feature selection

The features initially considered to be explanatory of a data-shift were further assessed for correlation with each other to identify redundancy and relevance in terms of distribution to identify features with little variance.

Figure 3.5 illustrates the correlation matrix of the feature set. According to Figure 3.5, the darker and larger blue bubbles represent a stronger positive correlation. And conversely, a darker and larger red bubble represents a stronger negative correlation.

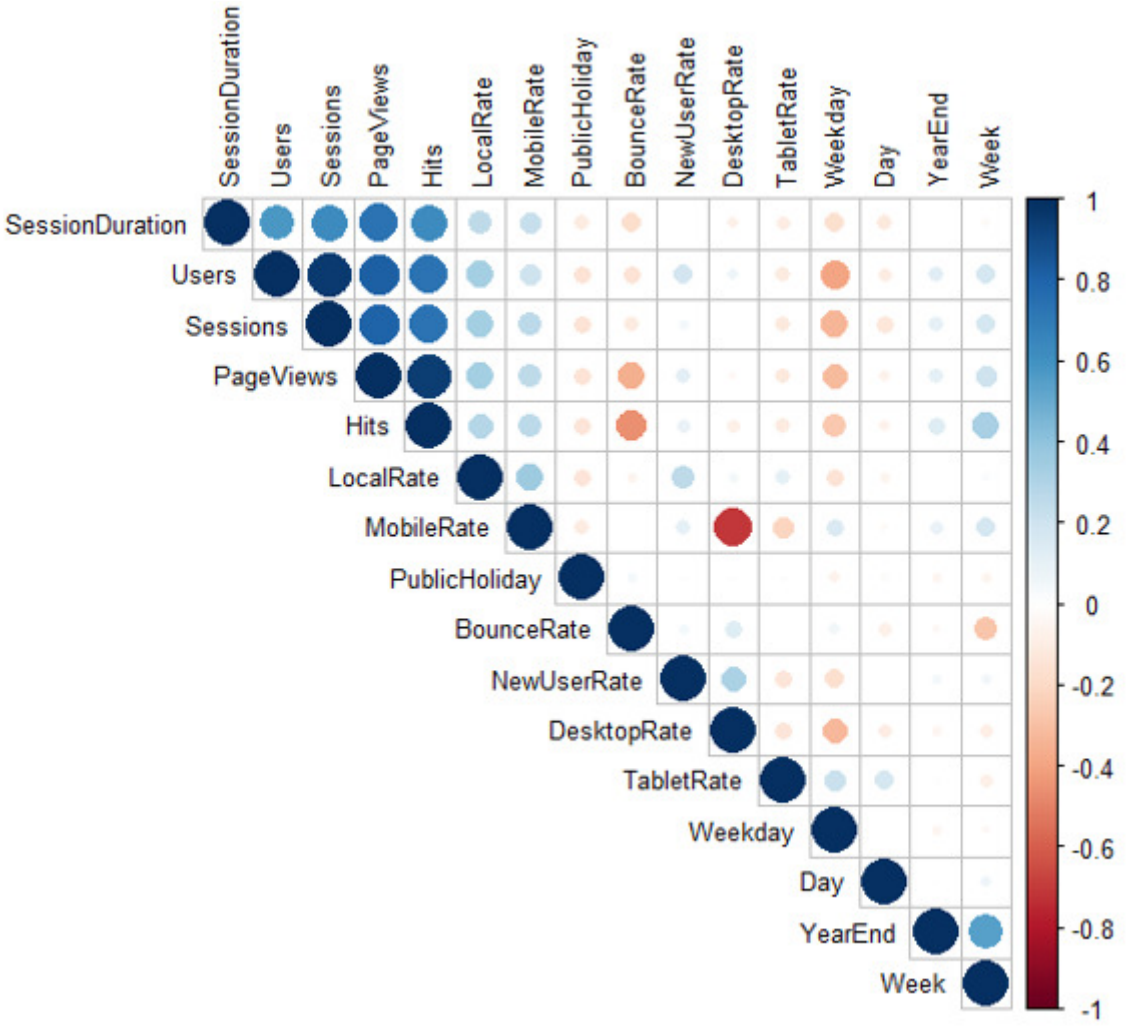

Figure 3.5: Correlation matrix between features considered for the ANN.

The correlation matrix indicated that the *hits* and *pageviews* features have shared a strong positive correlation and hence one had to be omitted. Similarly, the *users* and *sessions* shared a strong positive relationship as naturally, the user volume would influence the session volumes and due to the strength of the correlation, one ought to be omitted. Although the *sessionduration*-to-*pageviews* and *sessionduration*-to-*hits* features share a somewhat high positive correlation, it is possible that this may not always be the case. For instance, there may be users with a high volume of pageviews, yet a low session duration in cases where the users rushed through the website. Hence, to avoid possible loss of information, no omission would be implemented in this regard. A strong negative correlation was detected between the *mobilerate* and *desktoprate*. This was primarily driven by mobile and desktop devices being used frequently to access the website (tablet devices recorded a far lower incidence). Hence, to avoid possible loss of information from the tablet user-behaviour, both the *mobilerate* and the *desktoprate* were included within the model. Correlated features are known to hold redundant information and are known to potentially influence model performance negatively (Kavzoglu and Mather, 2002).

To identify features with no variance, Figure 3.6 depicts the box-and-whisker plots. Features with no variance ought to be omitted from the artificial neural model as such features would be irrelevant explanatory features (Kavzoglu and Mather, 2002).

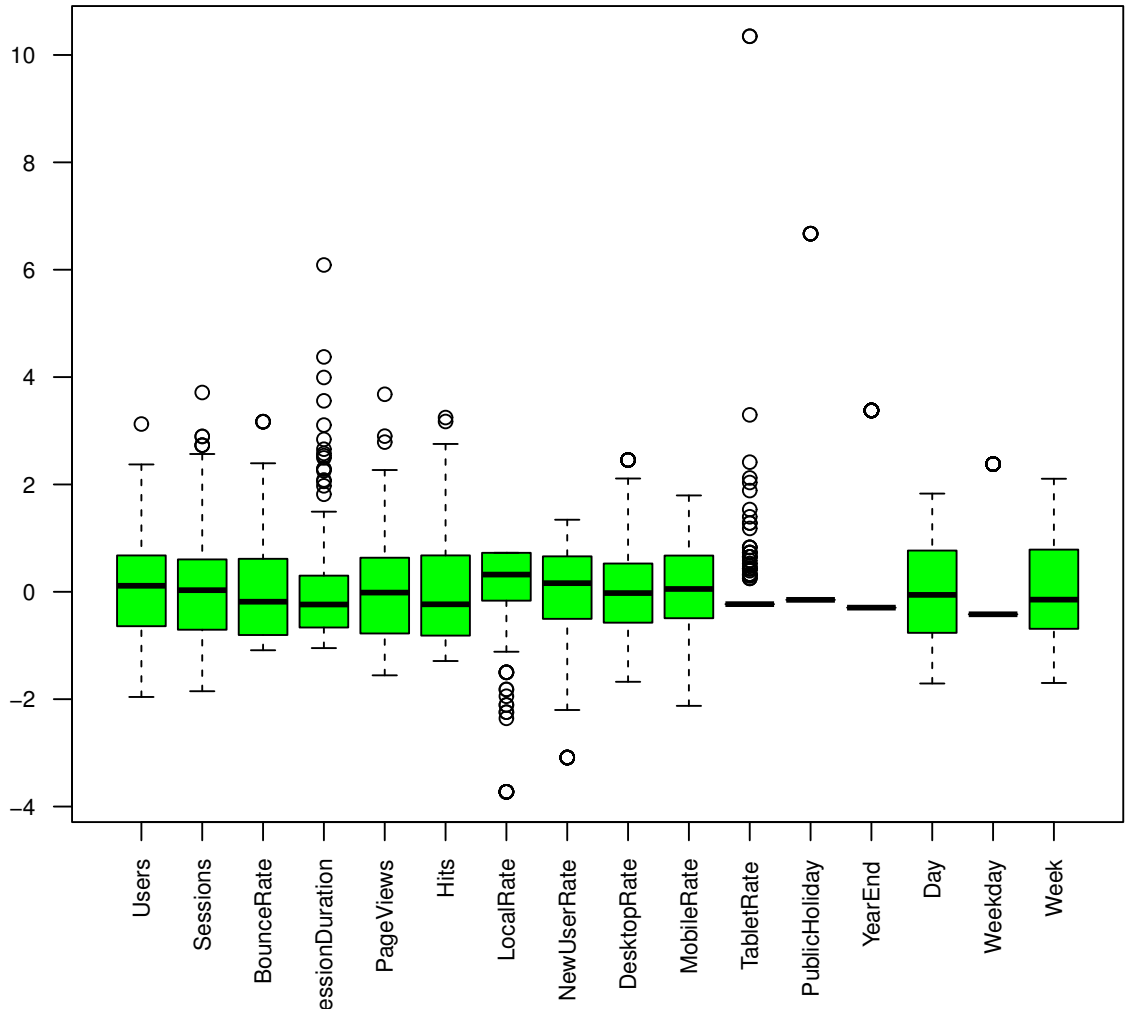

Figure 3.6: Box plots of scaled features considered for the ANN.

Most features considered to be included within the ANN model do show variation. The *publicholiday*, *yearend* and *weekday* features were binary flags and thus naturally did not hold much distribution. However, the *tabletrate* feature which quantifies the proportion of web visits that were from a tablet device on a given day shows minimal variation due to the low-incidence of table devices and thus was removed from the model.

Therefore, from the correlation and distribution assessments completed, the *hits*, *users* and *tabletrate* features were omitted from the ANN model.

3.3.3 Artificial neural net

This section discusses the artificial neural network constructed to detect a data-shift in the observed data. The proposed model could further be deployed to act as a sensor should a data-shift occur on web traffic data in the future. Although the model has been trained on the COVID-19 induced data-shift, it could potentially alarm when data volumes and expected behaviour change drastically which would inherently affect business profitability. The features identified for model building in Section 3.3.2 have been normalised prior to modelling (Van Gassen et al., 2020).

The observed data was randomly split into two subsets, a training dataset (80%) and a test dataset (20%). The training dataset was used to build the artificial neural network and the test dataset was used to validate the model accuracy.

The architecture of the network where eleven input layers (implicit and explicit features sourced from Google Analytics tracking). By design, there was one hidden layer with six nodes and the output layer. The output layer was the dependent variable that flagged the days that a data-shift occurred (Figure 3.7).

To assess the complexity and interpretability of the ANN, a model of eleven input layers, one hidden layer with six nodes and a single output layer was developed. Therefore, the total number of weights were $(11 * 6) + (6 * 1) = 72$. The number of biases were $6 + 1 = 7$. Thus, the total number of internal parameters equated to $72 + 7 = 79$. According to literature, the general principle suggested that the sample size ought to be roughly more than ten times the number of internal parameters to successfully train an ANN (Belhaj et al., 2021). This implied, a minimum sample size of 790 data points was required for a model of this complexity. Within this study, a base size of over 2000 data points were used to train the data-shift ANN.

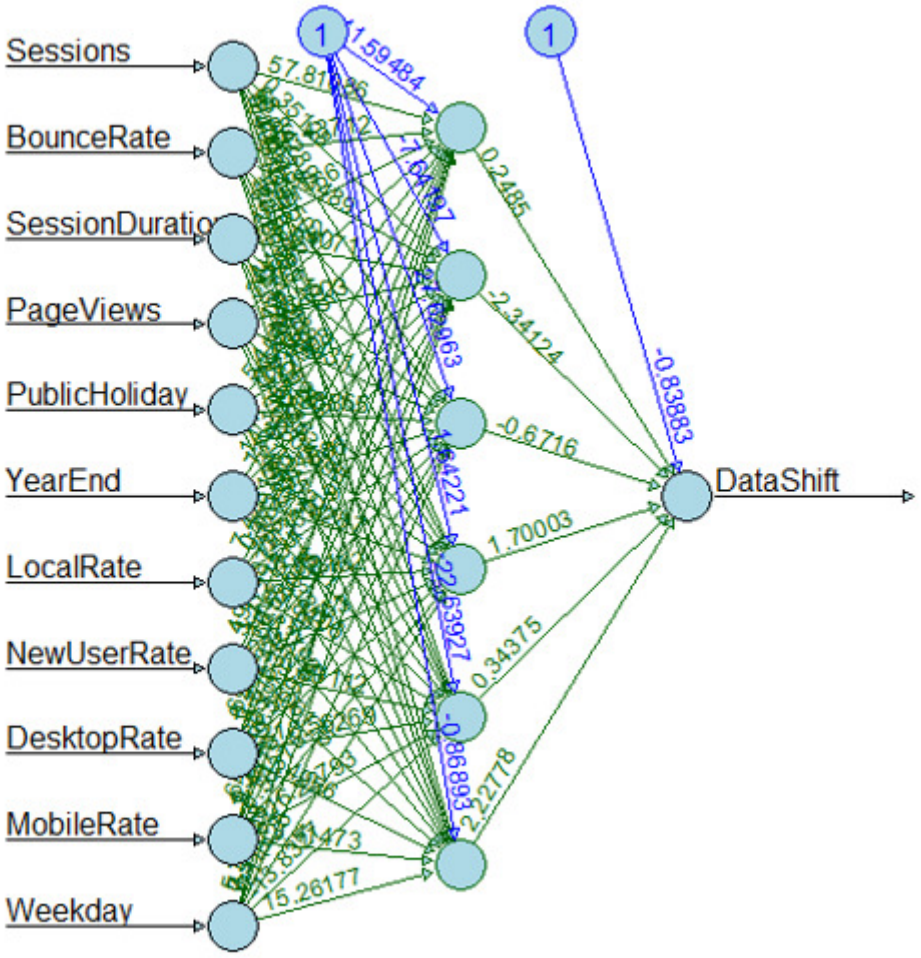

Figure 3.7: Constructed artificial neural network.

The dependent variable (datashift) naturally contained a class imbalance as the majority of the data points represented usual behaviour (datashift = 0). This was addressed through random under-sampling to maintain a proportion of 1:2 data-shift events to regular events prior to training the ANN models (Johnson and Khoshgoftaar, 2019). The purpose of the under-sampling was to avoid a majority bias that would affect the prediction results. The model fitted is illustrated in Equation 3.8:

$$\begin{aligned} DataShift \sim & sessions + bouncerate + sessionduration + \\ & pageviews + publicholiday + yearend + localrate + \\ & newuserrate + desktoprate + mobilerate + weekday. \end{aligned} \tag{3.8}$$

During the validation step, the model generated from the training dataset was validated against unseen data (the test dataset). In doing so, the model has yielded a 91.07% classification accuracy.

3.3.4 Feature importance

With the artificial neural network yielding satisfactory results, the feature importance's are discussed within this section and presented in Figure 3.8.

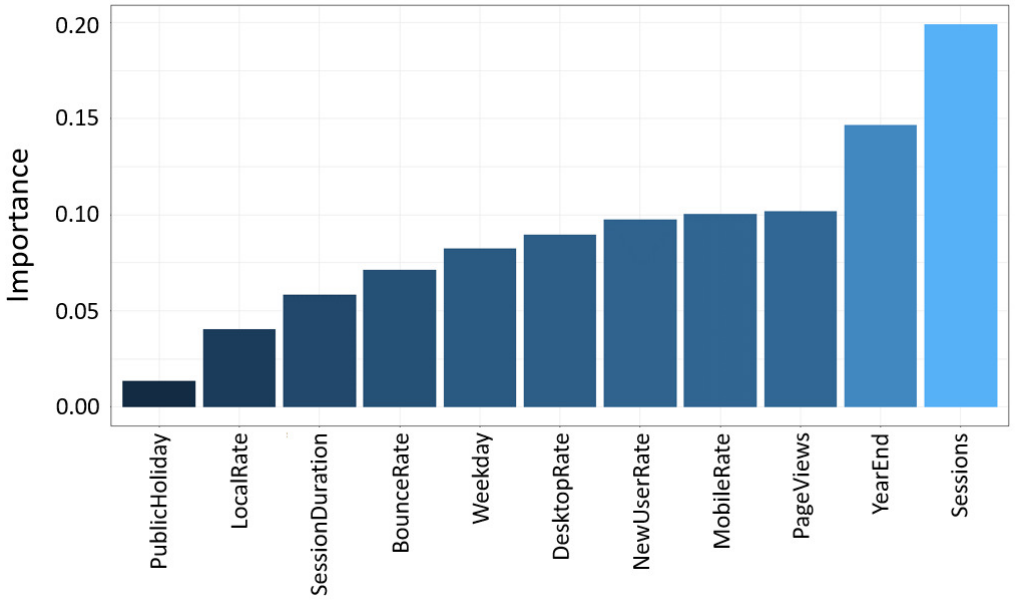

Figure 3.8: Feature importance in identifying an online web data-shift.

Figure 3.8 illustrates the feature importance scores that emerged whilst building the model. Each feature is assigned a percentage contribution so that the sum of all features included within the model sums to one hundred percent. Thus, the higher a feature's importance percentage, the more important the feature has proven to be. According to Figure 3.8, the most important features (represented by the tallest bars) are the volume of sessions, the year-end indicator, and the volume of pageviews. The volume of sessions and pageviews that talk to the traffic flow levels on a particular day. The *yearend* feature is merely a flag to indicate if a data point was during the December holiday period. The important features, thereby indicate that the model was able to accurately quantify the likelihood of a data-shift occurring by assessing the volume-based features (*sessions* and *pageviews*) keeping in mind the

time of year. The studied website was prone to seasonal patterns, and traffic would, understandably, dip during the December period. Thereby, the *yearend* feature was used to inform the model on when data volumes are expected to drop and not be considered a data-shift.
Whilst ANNs may be prone to overfitting, this study minimized the occurrence of overfitting due to an adequate base size given the complexity of network structure employed within the study. A network structure of eleven input layers, one hidden layer with six nodes and a single output layer requires a minimum of 790 datapoints (since the number of internal parameters were calculated to be 79) whilst the model was adequately trained on over 2000 data points. Additionally to avoid overfitting, the class imbalance inherent within the dependent variable was adjusted and finally the model was evaluated on a test sample to assess the classification accuracy to further detect for overfitting.

3.4 Summary

In this study, an artificial neural network model was used to model the COVID-19 induced data-shift in online web traffic on a South African informative website. The artificial neural network model has yielded a high detection accuracy in the event of a data-shift occurring even when deciphering between weekends and seasonal holidays (that would naturally result in a change in data patterns). The high detection accuracy yielded by the artificial neural network model suggested that the model can be employed as a sensor to flag the occurrence of a future data-shift. Although the model has been trained on a COVID-19 induced data-shift, the model may be used to detect data-shifts that are driven by other causes having similar symptoms.

According to the observed data, the website experienced a drastic drop in volume-based readings but fairly stable behaviour for those who have entered the website. Going into the harsh lock-down levels (when the data-shift occurred), the observed number of visits and users has dropped significantly. However, for the portion of users that did access the website during the period of the data-shift, the observed user-behaviour proved to be stable relative to the period prior to the data-shift (for instance, stable session duration and pageview volumes).

The artificial neural network highlighted that the volume-based measures (e.g., sessions and pageviews) were important indicators of a data-shift occurring whilst bearing in mind the time of year. The time of year was an important indicator since volumes would naturally drop during seasonal periods. Due to the extent of the data-shift detected on the studied website, the evidence suggested that the key research objectives of this study be addressed on data after the data-shift normalised. Nonetheless, the key takeout of the data-shift was the proposed artificial neural network model that could be employed to detect the occurrence of future data-shifts.
The data dependency of the ANN model was satisfied by a sufficiently large base size of over 2000 datapoints given the network structure of eleven input layers, one hidden layer with six nodes and a single output layer. Within the context of the study, the dependent variable labelled if the datapoint was a data-shift or not.

With the data-shift understood within Chapter 3 and a model available to detect future data-shifts, Chapter 4 identifies the features to be selected for unsupervised machine learning to determine the prevalent browsing patterns within the web traffic data.

CHAPTER 4

UNSUPERVISED MACHINE LEARNING FEATURE SELECTION ON WEBSITE DATA

4.1 Introduction

Website traffic often comprises numerous users from several different devices performing a wide variety of tasks and of varying engagement levels. Thus, attempting to summarise web traffic activity is considered a highly complex task. A simple but important, yet very difficult question to answer is: what are people doing on my website? For instance, some users spend a few seconds on the website whilst others perhaps hours; some users visit the website once whilst others several times; users have several different entry points onto the website and follow unique page paths (Thushara and Vamanan, 2016). Therefore, due to this complexity, unsupervised machine learning techniques (clustering algorithms) would assist in further aggregating the data based on browsing patterns inherent within the data. Prior to unsupervised machine learning, the process of feature selection is described as an important yet challenging problem (Wang and Zhu, 2008).

The success of supervised and unsupervised machine learning models depends highly on the features used for data modelling. It has been proven, that the set of features chosen could either improve or reduce the performance of statistical models (Dy and Bordley, 2004). Feature selection would also determine the computational costs and run-time associated with training models. Furthermore, feature importance scores are often used to interpret models and thus including the appropriate features is necessary (Dy and

Bordley, 2004). The initial set of features considered to be included within the model depends on the available features, the model's intuition, and research on similar models and applications. The features considered could be sourced directly from the data or indirectly (inferred) obtained using the available information (Ferreira and Figueiredo, 2012). Several researchers have described methods of feature selection for unsupervised machine learning techniques, however, this study focuses specifically on methods of feature selection prior to the unsupervised modelling on web traffic data from a South African informative website. The literature discussed provides a few examples of previous research on unsupervised machine learning feature selection (UFS).

In an application on microarray data, Wang and Zhu (2008) proposed an unsupervised feature selection technique that separates data points into clusters, and based on cluster contribution, features were selected. The framework of the study conducted focused on penalised model-based clustering. The researchers have found that the proposed methods have efficiently removed the non-informative features.

Fraiman et al. (2008) proposed UFS procedures targeted at identifying noisy non-informative features and multicollinearity between features that are appropriate to the forward-backward clustering algorithm employed. The methods were based on a two-variable-selection process and conditional means. The researchers found that the proposed methods did not work well for high-dimensional data.

Fop and Murphy (2017) classified UFS for model-based clustering into broad groups: Bayesian, penalisation and model selection approaches which have been applied to mortality data. The researchers concluded that feature independence is crucial and the aim is to discard both redundant and uninformative features during the feature selection process prior to unsupervised modelling.

Maugis et al. (2009) proposed an UFS process that classifies each feature as a relevant clustering feature, an irrelevant clustering feature dependent on a part of a relevant feature or an irrelevant clustering feature independent of all relevant features. The selection technique of identifiability and consistency

proved to be established. The researchers assessed random waveform data to illustrate the proposed feature selection process.

Chormunge and Jena (2018) proposed the use of correlation assessment to aid in UFS of high dimension data. The study discussed first feature elimination through K-means clustering and thereafter identification of non-redundant features through correlation measures from each cluster. The researchers state the experiment results, using microarray data, yielded accuracy and efficacy using the proposed method.

Guerif (2008) proposed UFS through a combination of multiple rankings. The experiment data showed that the approach yielded effective and stable results.

The discussed literature details the methods and applications of feature selection conducted by a few previous researchers. However, none provide an in-depth exploration of the feature selection process on such potentially big, robust and detailed web traffic data (at the time of writing).

This chapter aims to provide a generalised methodology that could be used by any corporation hosting a similar website to determine the key features that construct online user-behaviour groups. Chapter 4 further discusses the methodology in Section 4.2 and the empirical results in Section 4.3.

4.2 Methodology

This section discusses the background theory on methods that can be utilised to select the key features for unsupervised machine learning models for the attainment of accurate online user-behaviour groups.

4.2.1 Feature relevance

During the feature selection process, the variability of the features within the consideration set needs to be assessed. Features with no variability ought to be removed and those with little variability need to be further analysed. For instance, suppose a feature (say, age) has only a single value (age = 32) across all observations within the dataset (in an extreme case). This would

imply that the feature age is non-discriminant enough to be included within the model. Features with low variability can be included if the observations that differ could potentially provide insight into the unsupervised learning model. However, data scientists may choose to omit features with relatively low variance levels (although risking a potential loss of information to the model) in the attempt to optimise runtime in certain applications of machine learning.

Variance

The variance provides a measurement of how far spread observations are from the mean. For a random variable X, Equation 4.1 formulates the variance, where μ_i is the expectation (E) of X_i:

$$Var(X) = E[(X_i - \mu_i)^2]. \tag{4.1}$$

Although the variance metric is relative to the unit of measure of a particular feature, the higher the variance metric, the more spread observations are within the feature. Features with variance = 0 indicate that the feature observations are all identical. The coefficient of variation (CoV) of a random variable X is computed as the square root of the variance (standard deviation denoted by σ_X) divided by the mean (μ_X) of a feature (Equation 4.2)

$$CoV(X) = \frac{\sigma_X}{\mu_X}. \tag{4.2}$$

When the coefficient of variation is less than 1, the feature is said to have very low variability between the observations with the feature (Brown, 1998). However, this rule of thumb (coefficient of variance less than 1) has shown to be unreliable for very small mean values. This is driven by the calculation of the metric, with the denominator being the mean value, thus the closer the mean value is to 0, the larger the coefficient of variance metric will be. In this study, integer-valued features with mean values less than 0.2 and a coefficient of variance greater than 2.5 can also be used to identify features that had little variance about an approximate zero mean.

Mean absolute difference

The mean absolute difference is a measure of statistical dispersion within a numeric feature. The mean absolute difference of a variable X_i, is computed as per Equation 4.3:

$$MAD(X_i) = \frac{1}{n}\sum_{j=1}^{n}|X_{ij} - \bar{X}_i|, \tag{4.3}$$

where j represents each observation within X_i. The mean absolute difference provides an indication of the spread of the observations from the mean. The larger the mean absolute difference, the greater the variability within a feature. A mean absolute difference of zero implies that all observations within the feature are identical. Although there was no supporting literature to define a cut-off point to identify features with very low variability, this study identified features that have a mean absolute difference within 5% of the mean to be a low variability feature.

Dispersion ratio

The dispersion ratio of a variable X_i represents the ratio between the arithmetic and the geometric mean of the variable as per Equation 4.4:

$$Dispersion\ Ratio(X_i) = \frac{\mu_i}{GM_i}, \tag{4.4}$$

where $\mu_i = \bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$ and $GM_i = (\prod_{j=1}^{n} X_{ij})^{\frac{1}{n}} = e^{[\frac{1}{n}\sum_{i=1}^{n}\ln a_i]}$.

The closer a dispersion ratio is to 1, the lower the dispersion between the observations within a feature. In theory, the arithmetic mean would be larger than the geometric mean. However, the geometric mean is impacted by zero or missing values. Hence, the geometric mean (in cases of zero values) is often computed by excluding the 0 or missing data points which would, in turn, alter the expected relationship between the arithmetic and geometric means (de la Cruz and Kreft, 2018). At the time of writing, there is no literature to advise on a dispersion ratio value that could be used as a benchmark to identify low variability features. Thus, this study isolated features with a dispersion ratio between 0.8 and 1.2 as potential low variability features where zero data points were omitted from the geometric mean computation. However, depending on the scale of the variables and the distribution of the

data, an appropriate low variability interval would need to be determined in other applications.

Coefficient of unalikeability

The coefficient of unalikeability (u) computes how frequently observations differ from each other within variable X_i. It is often used as a pseudo measure of variance for categorical features of n observations by comparing one observation x_i with another observation within the same variable x_j for $i, j \in n$ and $i \neq j$ (Equation 4.5)

$$u(X_i) = \frac{\sum_{i \neq j} c(x_i, x_j)}{n^2 - n}, \tag{4.5}$$

where

$$c(x_i, x_j) = \begin{cases} 1, & x_i \neq x_j \\ 0, & x_i \neq x_j \end{cases}.$$

The coefficient of unalikeability computes a measure between 0 and 1 where the measures closer to 1 indicate the data within X_i are more unalike (Kader and Perry, 2007). There is no supporting literature to advise on the point at which the coefficient of unalikeability indicates very low variability. As a result, this study employed a rule of thumb that features with a coefficient of unalikeability less than 0.2 should be considered as possible low variability features.

4.2.2 Association between features

Within the feature selection process, it is important to inspect the association between features. Although measures with high or significant association should be included, this process will highlight potential information redundancy and features with potential natural relationships (Ferreira and Figueiredo, 2012). Thus if features share a high association, further exploratory analytics is required to decipher if the association indicates redundancy or insight. If features are highly associated indicating redundancy then one of the two features should be omitted from the unsupervised machine learning models. Similar to variability, data scientists are often required

to produce models that are lightweight in terms of run-time. Thus, during the feature selection process, associated variables may also be omitted (however potential loss of valuable insight needs to be inspected first). It is also important to note, an association may or may not be driven by a causal relationship.

Correlation matrix

A correlation matrix can be used to establish the association between numeric features with each other. The correlation $\rho_{(X,Y)}$ between two random variables X and Y, where X has a mean values μ_X and standard deviation σ_X. Suppose Y has a mean value μ_Y and standard deviation σ_Y is computed as per Equation 4.6:

$$\rho_{(X,Y)} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}. \tag{4.6}$$

The correlation matrix provides a metric ranging from -1 to +1. Features that share a correlation close to -1 imply that these features share a strong opposite relationship (when one is high, the other is low and vice-versa). Similarly, any two features sharing a correlation close to +1, share a strong positive relationship (when one is high, the other is also high). Features sharing a correlation close to 0 imply that these features have no relationship with each other. Features with an absolute correlation of between 0.68 and 1 are considered to share a strong or high correlation. Although features that share an absolute correlation of between 0.9 and 1 are considered to be very highly correlated (Taylor, 1990).

Chi-squared test for independence

The chi-squared test for independence (χ^2) can be used to assess the association between categorical features with each other. Although not all dependent variables are a concern, this method will highlight possible redundancy. Suppose a random sample, having n observations, which are classified into k groups (that are mutually exclusive), having observed numbers x_i ($i = 1, 2, \ldots, k$). Taking p_i as the probability that an observation falls into the ith class and taking the expectation defined as m_i=np_i, the χ^2 statistic is

calculated as per Equation 4.7

$$\chi^2 = \sum_{i=1}^{k} \frac{{x_i}^2}{m_i} - n. \tag{4.7}$$

It is important to note, that if significant dependencies do occur, features need not be removed. Rather, the significant dependencies will highlight areas for further investigation. For instance, if features are naturally dependent, then one of the two features ought to be removed. For example, the features country and region (similarly, international and country) will naturally have a dependency and thus one ought to be removed. The null hypothesis that tests if features are independent follows a significance test where the null hypothesis is rejected if the p-value is less than or equal to the level of significance (Alpha) (Holt et al., 1980).

Box-and-whisker plot

A box and whisker plot can be used to visualise the association between numeric and categorical features with each other. This visual representation can inform on the possible level of redundancy as well as the variability between the two assessed features. A box and whisker plot, is a visual representation that depicts the minimum value, maximum value, 25th percentile, 50th percentile and 75th percentile of the numeric features spread across the levels within the categorical feature (Thirumalai et al., 2017). For a categorical feature X with k levels and numeric feature Y, a box-and-whisker plot will indicate a strong association if, within the k levels of X, the observations contain the same value of Y (max-min = 0) and the Y value differs for the k levels or groupings of the k levels of X. The box-and-whisker plots within the k categories have non-overlapping interquartile ranges.

4.3 Empirical results

This section discusses acceptable approaches to gauge the variability within features and the correlation between features across various data types.

4.3.1 Measure of variability

The variability of numeric features was quantified using the variance, mean absolute difference and dispersion ratio statistics as reported in Table 4.1. The non-numeric features were assessed using the coefficient of unalikeability as shown in Table 4.2.

Table 4.1: Mean and measures of variability within numeric features.

	Mean and measures of variability						
Numeric features	Mean	Variance	Coefficient of variance	Mean absolute difference	Mean* 95%	Mean* 105%	Dispersion ratio
Accreditations	**0.1419**	0.1698	2.904	0.2496	0.1348	0.149	0.1262
Apprenticeship	**0.1521**	0.1856	2.8314	0.2647	0.1445	0.1597	0.1368
Bounces	0.3966	0.2429	1.2428	0.48	0.3768	0.4164	0.3954
Contact-us	**0.1949**	0.2665	2.6492	0.3291	0.1851	0.2046	0.1672
Courses	1.0812	2.2372	1.3834	**1.0988**	1.0271	1.1353	0.6165
Customised-engineering-trading	**0.0966**	0.1369	3.8316	0.1785	0.0918	0.1014	0.0814
daysSinceLast Session	1.7043	64.2067	4.7017	2.9873	1.6191	1.7895	0.2996
Distance	11.3806	1253.9402	3.1115	17.0262	10.8116	11.9496	22.1475
Engineering-academic-studies	**0.0051**	0.0154	24.1764	0.0102	0.0049	0.0054	0.0032
Engineering-Trade	**0.0154**	0.0442	13.6725	0.0305	0.0146	0.0162	0.0115
Hits	3.8376	18.338	1.1159	2.9534	3.6457	4.0295	1.5478
Home	1.1325	**0.7857**	0.7827	0.5831	1.0759	1.1891	**0.8934**
OrganicSearches	0.4726	0.2512	1.0604	0.4993	0.449	0.4963	0.4721
Pageviews	3.8342	18.3455	1.1171	2.9538	3.6425	4.0259	1.549
sessionCount	2.1487	13.5279	1.7117	1.6609	2.0413	2.2562	1.4899
SessionDuration	235.5274	249615.2537	2.1213	288.9087	223.751	247.3037	1.4289
Short-courses-skilled-programmes	**0.0103**	0.0204	13.9343	0.0204	0.0097	0.0108	0.0088
Trade-test-arpl	0.2068	0.3969	3.0458	0.3596	0.1965	0.2172	0.1477
University-of-technology-uot	**0.1838**	0.3571	3.2521	0.3245	0.1746	0.1929	0.1324

The values in bold font highlight the features that were detected to be low variability features according to the respective metrics as discussed in Section 4.2.

Table 4.2: Measures of variability within categorical features.

Categorical features	Coefficient of unalikeability
International	**0.1420**
Country	**0.1473**
UserType	0.4167
Region	0.5018
Browser	0.5045

From Table 4.1 the coefficient of variance indicated that the *home* feature contained low variance. Furthermore, for features with near zero mean values, the coefficient of variance highlighted that the following features held very low variability: *accreditations*, *apprenticeships*, *contact-us*, *customised-engineering-trading*, *engineering-academic-studies*, *engineering-trade*, *short-courses-skilled-programmes* and *university-of-technology-uot*. Similarly, the *courses* feature had a mean absolute difference within a 5% interval from the mean and the dispersion ratio identified the *home* feature to be a low variance feature. With regards to the categorical features, the coefficient of unalikeability (Table 4.2) indicated that the *international* and *country* features show little variability between observations. Exploratory analysis explained that most website visits are primarily from South Africa and a minimal portion of all visits are from elsewhere thus the low variability within the *international* and *country* features.

4.3.2 Measure of association

To assess the measures of association between features, the employed methods were a correlation matrix, a chi-squared test for independence and box-and-whisker plots.

Numeric to numeric features

The correlation matrix illustrated in Figure 4.1 expressed the association between the numeric features with each other.

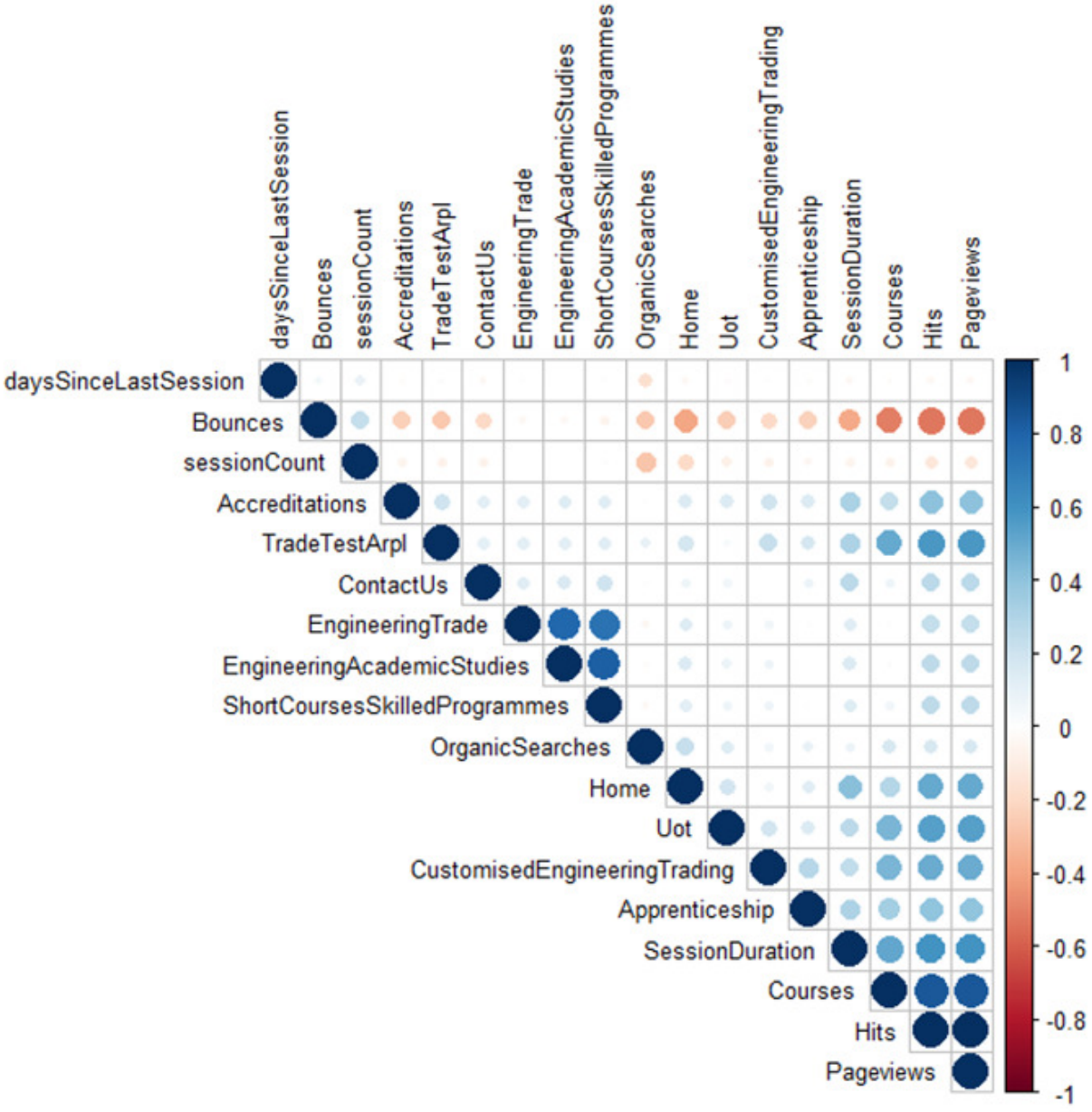

Figure 4.1: Correlation matrix of numeric features prior to unsupervised machine learning.

It is observed that a strong positive correlation exists between visits to certain webpages (*engineering-trade*, *engineering-academic-studies* and *short-courses-skilled-programmes*). Although this suggests a strong positive, it is not advised for such features to be removed due to the correlation as this relationship may be insightful.

A somewhat strong negative correlation appears between the *bounces* feature

and several other numeric features. Naturally, the higher the bounce rate, the less the interaction with the website. Since the event of a user bouncing is an important behaviour to monitor, the bounce feature was included within the unsupervised machine learning model.

However, the *hits* and *pageviews* features show to be very highly associated and to avoid redundancy, one of these two features should be omitted from unsupervised machine learning models. This was driven by the case study website, by design not encompassing much engagement per page with users. Thus, maintaining a *pageview*-to-*hit* ratio of 1:1.

Numeric to unordered categorical features

To inspect the association between the numeric features and the categorical features, box-and-whisker plots were constructed. Figure 4.2 depicts a few of the box-and-whisker plots between the categorical features and numeric features within the study. Table 4.3 labels the categorical features that have shown to have a strong association with the numeric features.

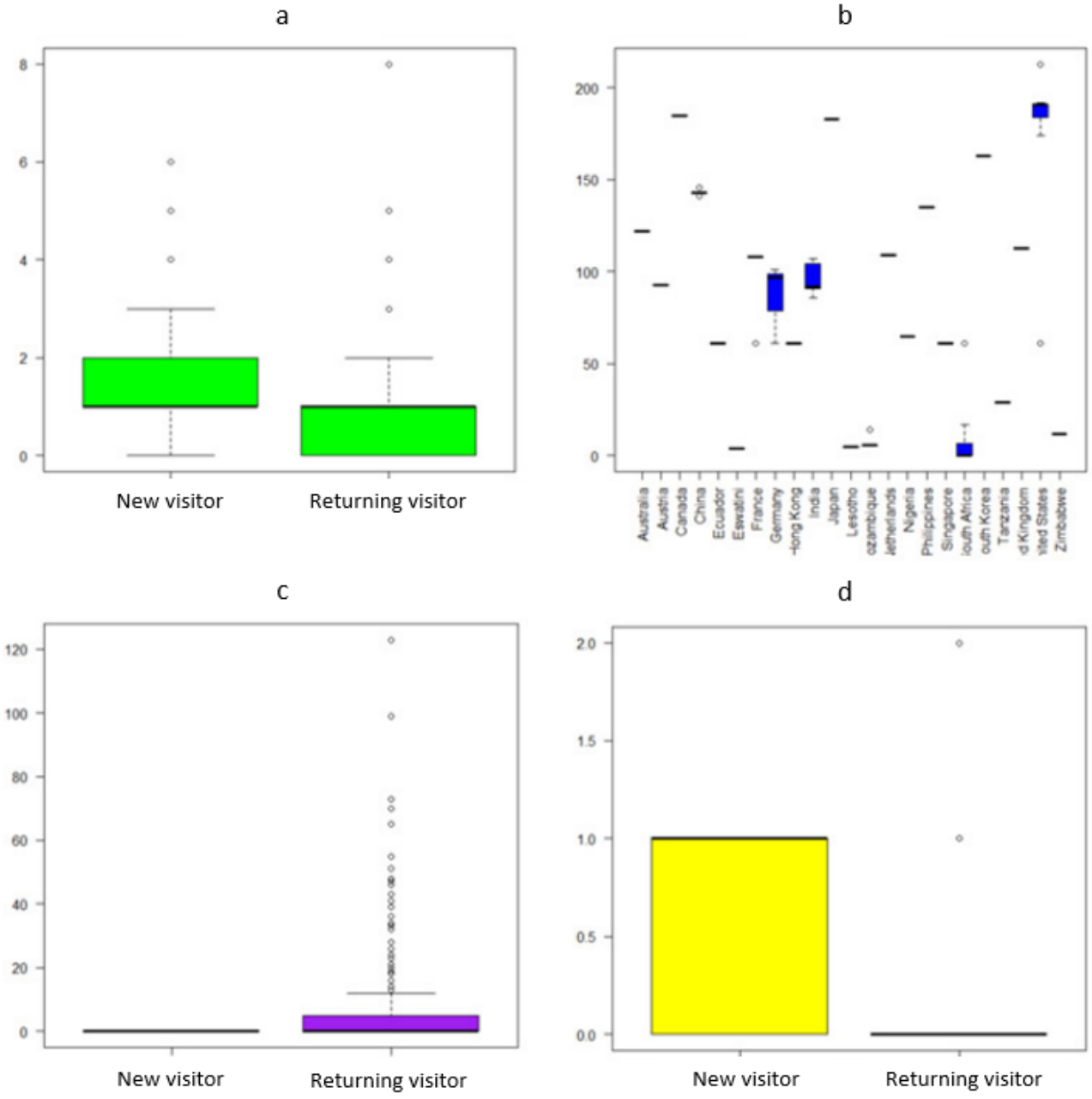

Figure 4.2: a) user type x home, b) country x distance, c) user type x dayssincelastsession, d) user type x organic searches.

Table 4.3: Box-and-whisker plot levels of associations.

	Categorical features			
Numeric features	**Browser**	**Country**	**Devicecategory**	**Usertype**
Accreditations	Low	Low	Low	Low
Apprenticeship	Low	Low	Low	Low
Bounces	Low	Low	Low	Low
Contact-us	Low	Low	Low	Low
Courses	Low	Low	Low	Low
Customised-engineering-trading	Low	Low	Low	Low
daysSinceLastSession	Low	Low	Low	High
Distance	Low	High	Low	Low
Engineering-academic-studies	Low	Low	Low	Low
Engineering-Trade	Low	Low	Low	Low
Hits	Low	Low	Low	Low
Home	Low	Low	Low	High
OrganicSearches	Low	Low	Low	High
Pageviews	Low	Low	Low	Low
sessionCount	Low	Low	Low	High
SessionDuration	Low	Low	Low	Low
Short-courses-skilled-programmes	Low	Low	Low	Low
Trade-test-arpl	Low	Low	Low	Low
University-of-technology-uot	Low	Low	Low	Low

The box-and-whisker plots illustrated in Figure 4.2 highlight potential measures of high association. The high association found between the features *usertype*-to-*dayssincelastsession*, *usertype*-to-*sessioncount*; and *country*-to-*distance* can be attributed to the natural relationships between these features. For features with natural relationships, one of the two features ought to be omitted due to redundancy of information. The numeric features are often chosen over categorical ones to avoid possible loss of information. Thus, *usertype* should not be included within the same model as *dayssincelastsession* and *sessioncount*.

Although the *country* and *distance* features have shown to have a naturally strong relationship, both these features may be included within the unsupervised model due to potentially insightful variation that occurs within a country. For instance, South Africa has a wider distribution of distances that users access the website from. Furthermore, a high association has been de-

tected between the features *usertype*-to-*home*; and *usertype*-to-*organicsearches*. This was driven by the tendency of new visitors viewing the *home-page* of the website more often than returning visitors. Returning visitors sought specific information on the website upon return. Returning visitors have shown to rarely search organically although this could be due to the device browser's capability to conveniently route to sites previously visited.

Categorical to categorical features

Chi-squared tests for independence have been used to establish the association between categorical features with each other. Table 4.4 presents the p-value measures of the chi-squared test for independence between the categorical variables.

Table 4.4: chi-squared test for independence: p-values.

Chi-squared test for independence	Browser	Country	Devicecategory	International	Region	Usertype
Browser		<0.0001*	<0.0001*	<0.0001*	<0.0001*	0.0036*
Country	<0.0001*		0.0109*	illogical	illogical	0.1921
Devicecategory	<0.0001*	0.0109*		<0.0001*	0.0094*	0.0708
International	<0.0001*	illogical	<0.0001*		illogical	0.0001*
Region	<0.0001*	illogical	0.0094*	illogical		0.4594
Usertype	0.0036*	0.1921	0.0708	0.0001*	0.4594	

Several variables show significant dependency on each other ($\alpha = 0.05$). Taking for instance the *browser* and *country* features, the dependency here is due to certain web browsers being more widely used within certain countries and since this is insightful, these measures should not be omitted from the unsupervised machine learning model due to this significant dependency. In cases where runtime is a concern, perhaps such omission of such features could be considered. The feature *usertype* (new or returning visitor) was shown to be the least dependent variable which only showed relation with the *international* feature. The *browser* and *usertype* features share a dependency due to returning visitors being primarily from South Africa, and South Africans most often use Google Chrome as the device browser of choice. However, of the features with significant dependencies identified, such relationships have shown to be insightful and thus none will be omitted due to these relationships.

Thereby, of the 25 features considered, for reasons of low variability and high association (e.g., natural relationships), 13 features are considered for omission from unsupervised machine learning models (a 52% degree of reduction in this study). Of which, there was one binary feature (*international*), two categorical (*region*, *usertype*) and ten numeric features (*accreditations*, *apprenticeship*, *contact-us*, *courses*, *customised-engineering-trading*, *engineering-academic-studies*, *engineering-trade*, *hits*, *short-courses-skilled-programmes*, *university-of-technology-uot*).

4.4 Discussion

In Chapter 4, methods for feature selection prior to unsupervised machine learning models on web traffic data have been explored. The evaluated methods focused on two important concepts: the variability of the features and the association between the features. The features considered within the study were of various data types and an appropriate method had to be applied accordingly. Using the metrics variance, mean absolute difference and dispersion ratio indicated that the features *accreditations*, *apprenticeships*, *contact-us*, *courses*, *customised-engineering-trading*, *engineering-academic-studies*, *engineering-trade*, *short-courses-skilled-programmes* and *university-of-technology-uot* should be omitted from an unsupervised machine learning model on account of low variability.

Although the *home* feature was also detected to contain low variability, the feature had insightful relationships with other features when the feature did vary. As discovered by the box-and-whisker plots and chi-squared tests, the features *user type*, *international* and *region* ought to be excluded due to natural relationships which would result in redundancy. Furthermore, the *hits* feature shared a strong positive correlation with *pageviews* and thus to eliminate redundancy the *hits* feature should be omitted. Although the features *engineering-academic-studies*, *engineering-trade* and *short-courses-skilled-programmes* have a fairly strong positive correlation despite having low variability. This suggests that although these pages were rarely viewed, these pages were often viewed together.

The outcome of this chapter is of tremendous value to data scientists and

corporates building online behaviour models. Such models are on the rise as the digital market continues to expand globally. However, much of the study would further contribute to unsupervised machine learning feature selection across several different applications.

Using the features identified for selection within Chapter 4, Chapter 5 constructs the unsupervised machine learning models to identify the prevalent intentions that emerge from the web traffic data studied.

CHAPTER 5

IDENTIFYING THE UNDERLYING INTENTIONS OF WEB VISITORS

5.1 Introduction

Data science techniques are often employed to make sense of the big and complex data (Shi, 2022). With a high volume of visitors each entering the website with different intentions and following unique page paths, the data at face value proved highly complex. To better understand the website usage in a manner that was reasonably ingestible, the study employed three unsupervised machine learning models to profile the web traffic data and determine the underlying intentions.

The underlying intentions were determined through the use of the K-means, hierarchical and DBScan unsupervised machine learning models. Although these models are considered traditional or classical, these models are nonetheless very popular (Cao et al., 2023). Apart from their popularity, these models were selected since each belongs to a different family of clustering methods: K-means belongs to the centroid-based or partition family, hierarchical clustering belongs to the connectivity-based family and the DBScan method belongs to the density-based family of clustering. In the attempt to determine the underlying intentions behind the web visits, the outputs of these traditional models were evaluated and compared with each other.

Given the nature of web analytics, the audience and subsequent behaviour are highly influenced by many factors (e.g., geo-location, core business, aesthetics, user experience and business seasonality (Uli and Laksmidewi, 2023)), thereby further adding to the novelty of this study. Although much research

has been conducted on web analytics and web-visitor profiling, as detailed in Section 1.3, none of the recent literature provided any details of a visit intent clustering model.

5.2 Model description and methods

This study employed three popular unsupervised machine learning algorithms that belong to different families of clustering methods. The K-means method belongs to the centroid-based or partition family of clustering, hierarchical clustering belongs to the connectivity-based family of clustering and the DBScan method belongs to the density-based family of clustering. Table 5.1 lists the notations denoted within Section 5.2.

Table 5.1: Notations used within the section.

Notation	Description
k	a real-valued integer to represent the number of clusters.
x	represents a single data point of feature X.
ϵ	a predefined hyper-parameter that sets the distance between a point and its neighbours.
n	minimum number of points within the cluster.

5.2.1 K-means

The K-means algorithm follows an iterative process that partitions the dataset into k (pre-defined) unique subgroups (clusters) where each data point resides in one cluster alone. The algorithm strives to ensure that the data points within each cluster are as similar as possible (homogenous) whilst the clusters are as different as possible from each other cluster. The iterative process minimises the sum of square distance between each data point and the cluster centroid. The cluster centroid represents the arithmetic mean of all data points within that cluster. The lower the variation within cluster data points, the more homogenous the cluster would be (Sinaga and Yang, 2020).
The pseudo-code of the K-means algorithm follows the process below:

1. User specifies the number of clusters (k).

2. The algorithm determines the centroids by randomly selecting the k data points for the centroids without replacement.

3. Iterate until the is no further change to the centroids (the assignment of data points to clusters remains the same).

4. Calculate the sum of squared distance between the data points and all centroids to determine cluster membership.

5. Compute the cluster centroids by taking the average of all data points within each cluster.

With a set of observations (X_1, X_2, ..., X_n) where each observation represents a d-dimension real vector, the K-means clustering algorithm strives to partition the n observed data points into K sets (where $K \leq n$). The sets, $S = S_1, S_2, ..., S_k$ where the within-cluster sum of squares (i.e., variance) is minimised (Likas et al., 2003) as per Equation 5.1.

$$\underset{S}{argmin} \sum_{i=1}^{k} \sum_{x \in S_i} \| x - \mu_i \|^2 = \underset{S}{argmin} \sum_{i=1}^{k} |S_i| Var S_i, \tag{5.1}$$

where μ_i represents the centroid of all data points in S_i as per Equation 5.2:

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x, \tag{5.2}$$

and $|S_i|$ represents the size of S_i. Therefore Equation 5.1 represents the minimisation of the pairwise squared deviations of points within each k cluster in Equation 5.3.

$$\underset{S}{argmin} \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{x,y \in S_i} \| x - y \|^2, \tag{5.3}$$

where

$$|S_i| \sum_{x \in S_i} \| x - \mu_i \|^2 = \frac{1}{2} \sum_{x,y \in S_i} \| x - y \|^2 . \tag{5.4}$$

Given that the total variance is constant, the algorithm served to maximise the sum of squared deviations between the data points in different clusters.

5.2.2 Hierarchical clustering

The hierarchical clustering method initiates by treating each observation as an individual cluster. Thereafter, the following two steps are iterated:

1. identify the two most similar clusters
2. thereafter, merge these two most similar clusters within this iteration.

The process continues to loop until all data points are merged into one cluster. The dendrogram provides a visual of the hierarchical clustering process as illustrated in Figure 5.1.

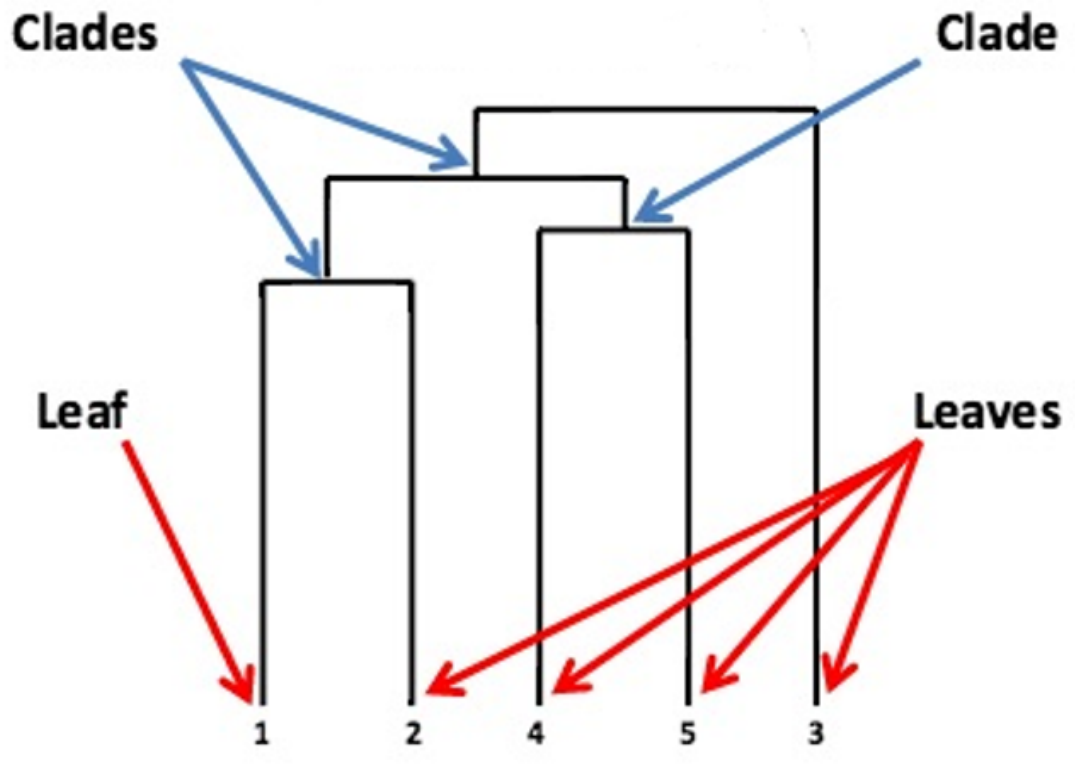

Figure 5.1: Sample dendrogram.

On a dendrogram, the clades represent the stacked branches. From the bottom-up, the clades that join together first would indicate the strongest similarity (Li et al., 2022).
Suppose $d(x, y)$ represents the distance between two data points x and y. There are three variations of agglomerative hierarchical clustering algorithms: single-linkage, complete-linkage and average-linkage clustering (Nielsen and Nielsen, 2016).

1. **Single-linkage clustering** represents the distance $D(X, Y)$ between any two clusters X and Y is computed as the minimum distance between data points within each cluster. As per Equation 5.5

$$D(X, Y) = \min_{x \in X, y \; in Y} d(x; y). \tag{5.5}$$

2. **Complete-linkage clustering** represents the distance $D(X, Y)$ between any two clusters X and Y is computed as the maximum distance between data points within each cluster. As per Equation 5.6

$$D(X, Y) = \max_{x \in X, y \ in Y} d(x; y). \tag{5.6}$$

3. **Average-linkage clustering** represents the distance $D(X, Y)$ between any two clusters X and Y is computed as the average distance between data points within each cluster. As per Equation 5.7

$$D(X, Y) = \frac{1}{|X||Y|} \max_{x \in X, y \ in Y} d(x; y). \tag{5.7}$$

Thereby, the hierarchical algorithm would merge the most similar clusters to create the hierarchy based on the type of distance metric employed within the algorithm (Li et al., 2022; Nielsen and Nielsen, 2016).

5.2.3 Density-based spatial clustering of applications with noise

The density-based spatial clustering of applications with noise (dbcsan) algorithm initiates by randomly selecting a single data point (x) and assigning it to cluster one. Then the algorithm assesses how many data points reside within the ϵ distance from x. If the count of such data points within ϵ distance is greater than or equal to the specified minimum number of points within the cluster (n), the algorithm would then consider this data point as a core point and will assign these ϵ-neighbours to the same cluster one. Thereafter, the algorithm examines each other member within cluster one and identifies their respective ϵ-neighbours. If any cluster-one members have n or more ϵ-neighbours, these data points will be added to cluster one. The process will continue growing cluster one until there are no more data points to add in (within ϵ distance and greater than or equal to the minimum number of data

points). The DBScan algorithm would then randomly pick another point from the dataset not belonging to any cluster and repeat the process. Data points that did not get assigned to any cluster will be labelled as outliers (Hanafi and Saadatfar, 2022).

Upon completion, each data point can be classified as one of the three:

- Core point: data point with at least the minimum number of neighbours within epsilon (ϵ) distance.
- Border point: data point with at least one core point within epsilon (ϵ) distance and less than the number of minimum neighbours within epsilon (ϵ) distance from itself.
- Noise point: data point with no core points within epsilon (ϵ) distance and thus could not be placed into a cluster.

DBScan definition and pseudo code

Definition 1. Let $N_\epsilon p$ denote the ϵ-neighbourhood of a point p be defined by $N_\epsilon p = \{q \in D | dist(p,q) \leq \epsilon\}$, where D represents a set of points and $dist(p,q)$ represents a distance function between two data points p and q.

Definition 2. A data point p represents a core point if $|N_\epsilon(p)| \geq minPts$.

Definition 3. A data point p is considered to be directly density-reachable from a point q with respect to ϵ and $minPts$ if $p \in N_\epsilon(q)$ and q is considered a core point.

Definition 4. A data point p is considered a border point if p is found to be directly density-reachable from a core point q and $|N_\epsilon(p)| < minPts$.

Definition 5. A data point p is considered to be density-reachable from a data point q with respect to ϵ and $minPts$ if there exists a chain of data points $p_1, ..., p_n$, with $p_1 = q$ and $p_n = p$ so that p_{i+1} is found to be directly density-reachable from p_i.

Definition 6. A data point p is considered to be density-connected to a data point q with respect to ϵ and $minPts$ if there exists a point o where both, p and q are density-reachable from o.

Definition 7. Let D represent a set of data points. A cluster C with respect to ϵ and $minPts$ is considered to be a non-empty subset of D satisfying the following conditions:

- $\forall\ p, q$: where $p \in C$ and q is considered density-reachable from p with respect to ϵ and $minPts$, then $q \in C$.
- $\forall\ p, q \in C$, p is considered to be density-connected to q with respect to ϵ and $minPts$.

Definition 8. A data point p is considered to be a noise if p is neither a core point nor a border point. In other words, a noise data point will not be grouped into any clusters.

Algorithm 1 details the DBScan clustering technique according to the definitions detailed above.

Algorithm 1 DBSCAN(D, ε, $minPts$)

1: {D is a set of unclassified points}
2: {ε is the maximum distance}
3: {$minPts$ is the minimum points to form a cluster}
4: Initialize cluster id $C = 0$
5: **for** each unclassified point $p \in D$ **do**
6: $N_\varepsilon(p) = RangeQuery(p, \varepsilon)$
7: **if** $|N_\varepsilon(p)| \geq minPts$ **then**
8: Set p's cluster id to C
9: $ExpandCluster(p, N_\varepsilon(p), C, \varepsilon, minPts)$
10: $C \leftarrow C + 1$
11: **else**
12: Label p as noise

Within Algorithm 1, the expand cluster step (at step nine) is further detailed in Algorithm 2. Algorithm 2 details the function $RangeQuery(p, \epsilon)$ represents all data points $q \in D$ and $dist(p, q) \leq \epsilon$

Algorithm 2 $ExpandCluster(p, NeighborPts, C, \varepsilon, minPts)$

```
for each point q ∈ NeighborPts do
    if q is unclassified then
        N_ε(q) = RangeQuery(q, ε)
        if |N_ε(p)| ≥ minPts then
            NeighborPts = NeighborPts ∪ N_ε(q)
    if q does not belong to any cluster then
        Set q's cluster id to C
```

At step one, the DBScan Algorithm 2 initiates with an arbitrary point p and retrieves all points in its ϵ-neighbourhood. If the count of data points in the ϵ-neighbourhood is sufficient (as defined by the specified $minPts$ parameter), a new cluster is started. This cluster is then expanded by adding in all its qualifying neighbours until it attains all data points which are density-reachable from p. However, if the count of data points in the ϵ-neighbourhood is insufficient, p will be marked as noise. At a later stage in the iteration, this data point might later be found in a sufficiently sized ϵ-environment of some other data point and hence fall into that cluster.

5.2.4 Cubic clustering criterion

Across clustering methods, determining the number of clusters remains an important aspect of unsupervised machine learning. The cubic clustering criterion (CCC) is a commonly used metric to give a sense of the acceptable number of clusters and potentially, the optimal number of clusters on a given dataset. The cubic clustering criteria quantify the deviation of the clusters from the expected distribution if all data points followed a uniform distribution. According to the cubic clustering criteria, larger positive values imply a better solution, as it indicates a larger variance from a uniform (no clusters) distribution. The cubic clustering criteria values greater than two indicate good clusters. Cubic clustering criteria values between zero and two indicate potential clusters but should be taken with caution. Furthermore, large negative values may indicate outliers. With the acceptable number of clusters in mind, the chosen number of clusters is often determined based on the profiling of the clusters in the context of the data in real-world applications (Joshi et al., 2022).

5.2.5 Silhouette coefficient

The silhouette coefficient is a commonly computed metric employed to assess the accuracy of clustering solutions. The silhouette coefficient ranges between -1 to +1 and implies the following (Bagirov et al., 2023):

- Clusters with silhouette coefficients closer to +1 imply very tight observations within the cluster (homogenous)
- Clusters with silhouette coefficients closer to 0 imply possible overlapping clusters
- Clusters with silhouette coefficients closer to -1 imply observations are not very similar.

5.2.6 Unsupervised modelling data preparation

Given the features selected for unsupervised machine learning methods in Chapter 4, the numeric features with a small degree of extreme value outliers were identified (Section 2.4.1) and approximated through a winsorisation clean to remove the influence of extreme values (Sullivan et al., 2021). Since most features have originated from the tracking tool, being system-generated metrics, there were minimal missing data points on the observed dataset (Section 2.4.2). Categorical features with numeric associates have been eliminated to reduce redundancy. For instance, the feature user type which labelled if a visitor was new or repeat could be determined from the days since the last session feature as detailed in Chapter 4. Similarly, the region and country information could be determined through the distance feature.

With the categorical features selected in Chapter 4, indicator variables were created using the one-hot encoding technique (Pargent et al., 2022). This transformation was necessary as the unsupervised machine learning models employed within this study required numeric inputs. For example, the mobile device brand feature was replaced with individual features for each of its brands so that a feature Samsung was created that contained a value of one if the brand of the device was Samsung and zero if any other brand was used.

Finally, the feature set was normalised to ensure a consistent scale prior to the unsupervised machine learning models (Van Gassen et al., 2020).

5.3 Empirical results

This section discusses the intents or clusters that were identified within the data by the three unsupervised machine learning methods employed.

5.3.1 Number of clusters

To gauge the acceptable number of segments (intents) that prevail within the studied data, the cubic clustering criterion (CCC) metric was computed. Figure 5.2 illustrates the CCC metrics between a two-cluster solution and a ten-cluster solution.

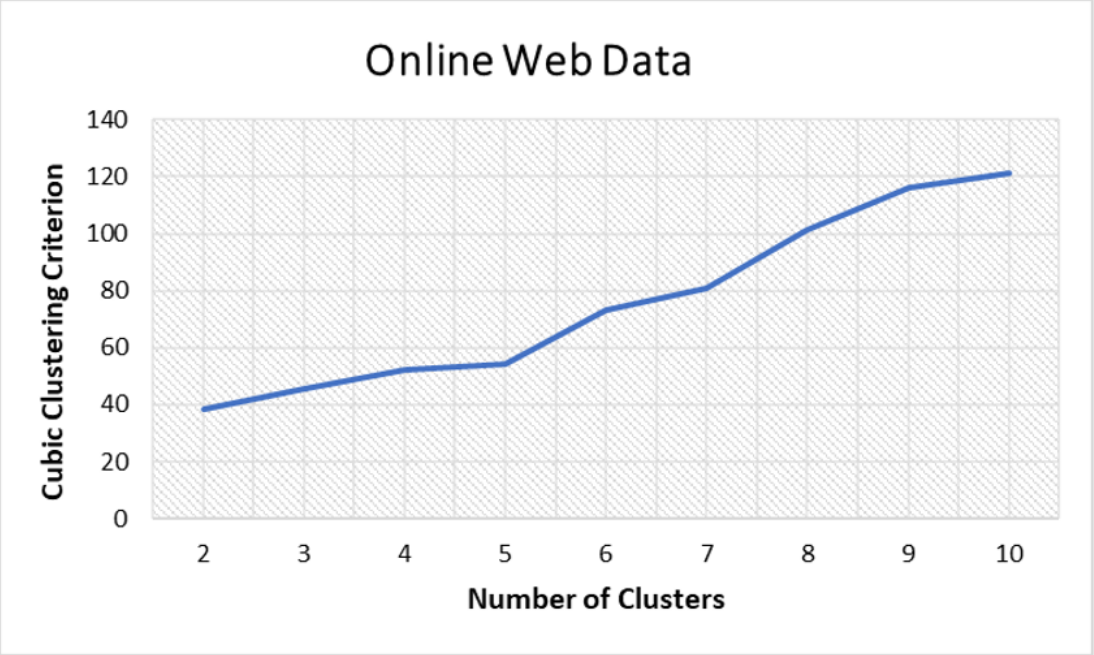

Figure 5.2: The cubic clustering criterion values.

According to the CCC measures, segmenting the data into between two to ten clusters will yield acceptable clustering solutions. However, the CCC metric also implied that the more clusters created, the better the clustering solution would be. During the study, across the several clustering methods, between three to eight clustering solutions were developed. It was found that

the five cluster solutions resulted in the most appropriate segments/intents in the context of this study across the three unsupervised machine learning methods applied. Using the clustering solutions that ranged between three to eight, each clustering solution was profiled by studying the cluster homogeneity across the input features (e.g., session duration or distance) and the corresponding cluster size to determine the optimal number of clusters. In doing so, the five-cluster solution proved to be the most insightful across the three clustering methods.

5.3.2 K-means

This section discusses the clusters that emerged from the K-means model. The K-means model was employed to segment the online web visit data to identify the underlying intents expressed across the visits. A range of clustering solutions was developed from a three-cluster solution to an eight-cluster solution. Upon profiling the data, the five-cluster solution yielded results that best explained the online web intents. Table 5.2 and Table 5.3 profile the K-means five-cluster solution against various online metrics.

Table 5.2: K-means cluster profiling across web metrics.

K-means clusters	Cluster size	% Size	Avg. sessionduration	Avg. bounces rate	Avg. organic search rate	Avg. hits	Avg. sessioncount	Avg. dayssincelastsession	Avg. distance	Avg. pageviews
1	988	15.0%	159.84	7.0%	55.0%	5.42	1.55	3.42	16.14	5.40
2	3260	49.6%	43.98	34.0%	40.0%	2.47	1.54	2.48	32.51	2.45
3	1066	16.2%	398.56	0.0%	66.0%	11.11	1.39	3.18	7.17	11.02
4	778	11.8%	196.44	7.0%	61.0%	8.41	1.38	2.97	7.35	8.36
5	477	7.3%	611.91	0.0%	70.0%	26.02	1.27	3.00	4.74	25.94

Table 5.3: K-means cluster profiling across webpage visits.

K-means clusters	Home page index	Accreditations page index	Apprenticeship page index	Contact-us page index	Courses page index	Engineering-Trade page index	Engineering-academic-studies page index	Customised-engineering-trading page index	Short-courses-skilled-programmes page index	Trade-test-arpl page index	University-of-technology-uot page index
1	0.94	0.09	0.08	1.15	0.26	0.02	0.06	0.02	0.00	0.11	0.04
2	0.86	0.05	0.05	0.00	0.24	0.02	0.04	0.01	0.00	0.05	0.03
3	1.52	0.22	0.66	0.11	1.98	0.37	0.11	0.20	0.00	0.66	0.26
4	1.15	0.11	0.11	0.15	0.97	0.16	0.17	0.09	1.21	0.20	0.10
5	1.95	0.64	0.79	0.45	2.10	0.82	1.07	0.92	1.18	1.17	0.82

Table 5.4: K-means cluster summary.

K-means clusters	Silhouette coefficient	Outstanding attributes
1	0.31	Over-index on the *contact-us* page, low bounce rate, moderate visit duration.
2	0.40	Very low session duration, very low pageviews, high bounce rate, furthest distance from the corporate location.
3	-0.05	Fairly high session duration, fairly high pageviews, low bounce rate.
4	0.13	Over-index on the *courses* page and *short-courses* page, moderate session duration, low bounce rate.
5	0.01	Very high session duration, very high pageviews, very close geo-proximity to the corporate coordinates, low bounce rate and high organic search rate.

Table 5.2 presents key online metrics across the five clusters and Table 5.3 indicates the average number of times that cluster members visited each webpage. This thereby allowed the clusters to be further understood in terms of the similarity of visits within each cluster and the differences between the clusters themselves. According to the K-means model, cluster one resembled the group of visits that intended primarily to contact the corporation. During cluster one visits, the bounce rate was low, the visits were relatively in close geographic proximity to the corporate and all visitors had primary interest in the *contact-us* page. Cluster two contained visits that expressed very little interest in the website. Cluster two indicated visits that would either bounce or shortly drop-off the website. Furthermore, in cluster two, visitors' geo-location was on average the furthest away from the corporate's coordinates. Although cluster two proved to be the largest cluster identified by the K-means model, it represented a group of visits that were of very low engagement. According to the model, cluster three represented a group of visits that were fairly engrossed with a moderate engagement with the web-

site. Cluster four represented a group of visits that were seeking specific information and dropped-off thereafter. It was evident that cluster four had a particular interest in the *courses* and *short-courses* webpages of the website. This page detailed the courses that the corporation had to offer. Cluster five represented a group of visits that were very engrossed and had a high engagement with the website. Within cluster five, the session duration, volume of pageviews and the rate of organic searching (which indicated that visits were not routed to the website through a link) were relatively the highest. Table 5.4 presents the average cluster silhouette scores, with clusters one and two indicating strong homogeneity. Clusters three and five grouped visits that were very engaged on the website and thus contained a wide variety of activities expressed resulting in poorer silhouette scores. Cluster four expressed a moderate silhouette score. Given the understanding behind the lower silhouette scores on clusters three and five, the K-means model yielded satisfactory results given the real-world data.

5.3.3 Hierarchical clustering

The second unsupervised machine learning model employed a hierarchical clustering method. The subsequent dendrogram of the hierarchical model is presented in Figure 5.3.

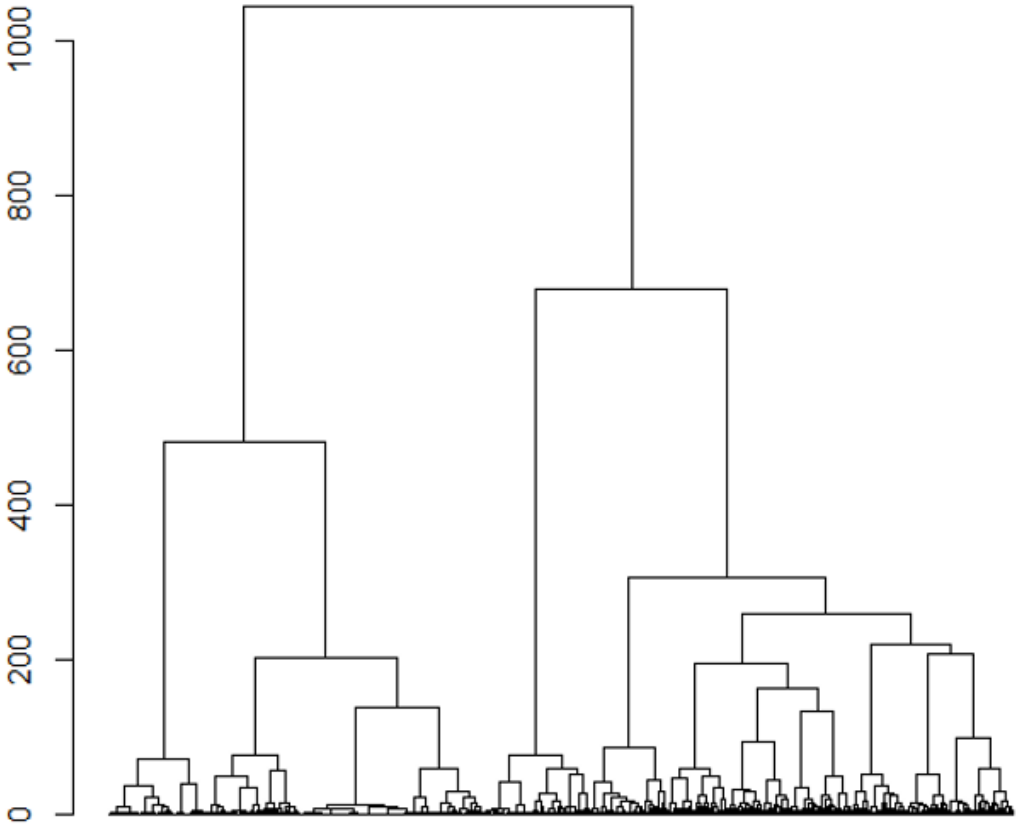

Figure 5.3: Online web visit dendrogram.

Upon profiling the clustering solutions, the five-cluster solution proved to be most appropriate in the context of this study. Table 5.5 and Table 5.6 illustrate the cluster profiling of the hierarchical clustering model.

Table 5.5: Hierarchical cluster profiling across web metrics.

Hierarchical clusters	Cluster size	% Size	Avg. sessionduration	Avg. bounces rate	Avg. organic search rate	Avg. hits	Avg. sessioncount	Avg. dayssincelastsession	Avg. distance	Avg. pageviews
1	2509	38.2%	306.63	6.0%	57.0%	11.67	1.63	6.02	8.26	11.60
2	551	8.4%	202.87	9.0%	62.0%	7.96	1.35	1.24	8.44	7.88
3	720	11.0%	4.06	56.0%	22.0%	1.54	1.03	0.02	124.30	1.54
4	1994	30.4%	91.47	29.0%	51.0%	2.92	1.48	0.15	4.30	2.90
5	795	12.1%	131.41	9.0%	53.0%	4.53	1.50	3.14	17.77	4.52

Table 5.6: Hierarchical cluster profiling across webpage visits.

Hierarchical clusters	Home page index	Accreditations page index	Apprenticeship page index	Contact-us page index	Courses page index	Engineering-Trade page index	Engineering-academic-studies page index	Customised-engineering-trading page index	Short-courses-skilled-programmes page index	Trade-test-arpl page index	University-of-technology-uot page index
1	1.28	0.35	0.53	0.22	1.30	0.38	0.33	0.30	0.34	0.49	0.35
2	1.15	0.00	0.14	0.20	0.96	0.03	0.21	0.05	1.19	0.18	0.02
3	0.68	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.00
4	1.07	0.00	0.00	0.00	0.46	0.00	0.00	0.00	0.00	0.14	0.00
5	0.89	0.00	0.00	1.15	0.21	0.00	0.00	0.00	0.00	0.09	0.00

Table 5.7: Hierarchical cluster summary.

Hierarchical clusters	**Silhouette coefficient**	**Outstanding attributes**
1	-0.21	High session duration, high pageviews, low bounce rate.
2	0.15	Over-index on the *short-courses* page and *courses* page, moderate session duration, low bounce rate, highest organic search rate.
3	0.60	Very low session duration, very low pageviews, high bounce rate, furthest distance from the corporate location.
4	0.38	Moderate session duration, brief pageviews, moderate bounce rate, close proximity to corporate coordinates.
5	0.39	Over-index on the *contact-us* page, low bounce rate, moderate session duration.

According to the hierarchical clustering model, cluster one represented visits that were highly engaged with the website, with a high average session duration, multiple pageviews, relatively close geographic proximity to the corporate's coordinates and a low bounce rate. Cluster two identified the individuals who were seeking specific information and thereafter left the website. Cluster two visits mainly viewed the *short-courses* and *course* pages. Cluster three, represented visits that resembled accidentally or with very little interest landing onto the website. Such visits were from the furthest geoproximity away from the corporation's offices (many of which were from countries outside of South Africa). Cluster three also recorded the highest bounce rate. Cluster four represented visits that showed evidence of disengagement due to dropping-off after spending a brief amount of time on the website. Cluster five represented the group of visits that were primarily interested in the contact information from the website and thereafter left the website. This can be inferred from the short visit duration but all visits index-

ing high on the *contact-us* webpage. The cluster silhouette coefficients of the hierarchical clustering model are recorded in Table 5.7. Clusters three, four and five attained strong silhouette scores implying good cluster homogeneity. Cluster two attained a moderate silhouette score. Cluster one recorded a low silhouette score due to the nature of the visits grouped. Within cluster one, visits that were very engaged with the website and expressed their high engagement in various ways (following unique paths and visiting pages in differing manners) and thus mathematically resulted in poor silhouette scores. Overall, the silhouette scores implied that the hierarchical clustering model yielded satisfactory results.

5.3.4 DBScan

The third unsupervised machine learning method applied was the DBScan model to identify the visit intents on the studied website. The five-cluster solution (inclusive of the noise-cluster) proved to be the most appropriate in the context of this study when employing the DBScan method. Table 5.8 and Table 5.9 further illustrate the cluster profiles per segment.

Table 5.8: DBScan cluster profiling across web metrics.

DBScan clusters	Cluster size	% Size	Avg. sessionduration	Avg. bounces rate	Avg. organic search rate	Avg. hits	Avg. sessioncount	Avg. dayssincelastsession	Avg. distance	Avg. pageviews
0	3151	48.0%	331.19	5.0%	56.0%	11.17	1.62	5.74	10.68	11.10
1	2440	37.1%	26.41	40.0%	42.0%	2.05	1.37	0.16	39.13	2.04
2	592	9.0%	58.72	8.0%	58.0%	3.78	1.34	0.15	11.27	3.77
3	270	4.1%	78.29	15.0%	60.0%	5.07	1.21	0.11	5.33	5.05
4	116	1.8%	60.01	3.0%	67.0%	3.44	1.30	0.13	2.87	3.38

Table 5.9: DBScan cluster profiling across webpage visits.

DBScan clusters	Home page index	Accreditations page index	Apprenticeship page index	Contact-us page index	Courses page index	Engineering-Trade page index	Engineering-academic-studies page index	Customised-engineering-trading page index	Short-courses-skilled-programmes page index	Trade-test-arpl page index	University-of-technology-uot page index
0	1.32	0.28	0.40	0.30	1.28	0.31	0.30	0.25	0.39	0.54	0.28
1	0.91	0.00	0.00	0.00	0.24	0.00	0.00	0.00	0.00	0.00	0.00
2	0.78	0.00	0.00	1.08	0.06	0.00	0.00	0.00	0.00	0.00	0.00
3	0.95	0.00	0.00	0.00	0.80	0.00	0.00	0.00	1.09	0.00	0.00
4	0.46	0.00	1.17	0.00	0.16	0.00	0.00	0.00	0.00	0.00	0.00

Table 5.10: DBScan cluster summary.

DBScan clusters	Silhouette coefficient	Outstanding attributes
0	-0.11	High visit duration, high pageviews, low bounce rate.
1	0.45	Very low session duration, very low pageviews, high bounce rate, furthest average distance from the corporate location.
2	0.45	Over-index on the *contact-us* page, low bounce rate, moderate session duration.
3	0.21	Over-index on the *short-course* page, high index on the *courses* page, moderate session duration, low bounce rate.
4	-0.34	Over-index on the *apprenticeship* page, moderate session duration, low bounce rate, high organic search rate.

Upon profiling the clusters of the DBScan model, cluster one represented visits that imply accidental landing on the webpage with very low engagement with the website. Cluster one maintained the highest bounce rate, minimal duration and the furthest geo-proximity with on average two pages viewed per visit. Cluster two represented visits that expressed main interest in the *contact-us* page alone. Cluster three comprised of visits that expressed primary interest in the *courses* and *short-courses* page whilst cluster four portrayed primary interest in the *apprenticeships* webpage. Across clusters two, three and four, visits were interested in specific webpages. It was also observed that cluster zero, which the model termed as the noisy data points, represented visits that were relatively the most engrossed, and spent the most amount of time on the webpage with a high volume of webpages visited. According to the cluster silhouette coefficients as tabulated in Table 5.10, the DBScan clustering model resulted in satisfactory results. Clusters one, two and three recorded high silhouette coefficients. Cluster one represented the high-engagement group of visits that naturally expressed a high variance in

behaviour patterns (visits spending a longer duration on the website, following unique page paths, and visiting different webpages). Cluster four recorded a poor silhouette score which could be attributed to the cluster size.

5.4 Final thoughts

This study sought to understand the intents behind the website visits of a South African informative website. In doing so, the study employed three unsupervised machine learning models, each belonging to a different clustering family, namely, the K-means, hierarchical and DBScan clustering methods. Across the three different clustering techniques, it was evident that five distinct visit intents were observed within the data. An intent described as *engrossed* was evident in the results. The *engrossed* segment represented the group of visits that were highly engaged with the website, spent a long duration on the website, visited multiple pages and were often within close proximity to the corporate's coordinates (Google Analytics tracking supplies the coordinates of the device whilst on the website). An intent described as *seekers* was discovered. The *seekers* represented a group of visits that went online, visited specific webpages, retrieved what was needed and thereafter exited the website (particularly on the *courses* and *short-course* pages). An intent described as *accidental* was discovered. The *accidentals* were visits that recorded a high bounce rate, just stepped into the website then hopped-off. It was noticed that the *accidentals* contained most of the foreign web visitors. An intent that can be described as the *drop-offs* was discovered. The *drop-offs* would enter the website, explore a few pages, and then afterwards drop-off with low to medium engagement on the website. And lastly, an intent that can be described as *get-in-touch* was discovered. The *get-in-touch* represented visits that had a primary interest in the *contact-us* webpage and thereafter immediately exited the website. Table 5.11 shares the intents that were discovered by three unsupervised machine learning models.

Table 5.11: Web visit cluster comparison.

Method	Cluster	Size	Silhouette coeff.	Persona
K-means	1	15.04%	0.31	Get-in-touch
	2	49.63%	0.40	Accidentals/Drop-offs
	3	16.23%	-0.05	Engrossed: Moderate engagement
	4	11.84%	0.13	Seekers
	5	7.26%	0.01	Engrossed: High engagement
Hierarchical	1	38.19%	-0.21	Engrossed
	2	8.39%	0.15	Seekers
	3	10.96%	0.60	Accidentals
	4	30.35%	0.38	Drop-offs
	5	12.10%	0.39	Get-in-touch
DBScan	0	47.97%	-0.11	Noise (resembles Engrossed)
	1	37.14%	0.45	Accidentals/Drop-offs
	2	9.01%	0.45	Get-in-touch
	3	4.11%	0.21	Seekers: Short-courses
	4	1.77%	-0.34	Seekers: Apprenticeships

The K-means method has identified all intents but merged the *accidentals* and *drop-offs* into one cluster. Furthermore, the K-means model further split the *engrossed* segment into moderate engaging visits and high engaging visits. The DBScan method was also able to identify the common intents but split the *seekers* into two small volume intents whilst merging the *accidentals* and drop-offs. Furthermore, the *engrossed* visits were all labelled as noise data points by the DBScan method. Upon comparison of the subsequent clusters generated by each of the three employed unsupervised machine learning methods, the hierarchical clustering displayed superior results. The hierarchical cluster had successfully isolated the five intents portrayed within the data, the cluster sizes were satisfactory, and the clusters maintained superior homogeneity as quantified by the silhouette coefficients relative to the K-means and DBScan models. Across the three unsupervised machine learning models employed, all models have performed exceptionally well in finding meaningful clusters within the data. However, in the context of this study, the hierarchical model yielded the results that best suited the research needs of this study on the observed data.

Across the unsupervised machine learning models, the volume of drop-off visits observed was concerning. Therefore, Chapter 6 employed survival

models to further identify the key hazards behind the volume of drop-offs.

CHAPTER 6

SURVIVAL MODELS TO IDENTIFY THE KEY HAZARDS OF WEBSITE DROP-OFFS

6.1 Introduction

On the studied website, roughly one in five visitors would enter the website and thereafter leave after viewing no more than three webpages of the website (after the exclusion of bounced visits). Like most corporates, TEKmation strives to further optimise the website to better meet customers' needs to ultimately increase market share (Awichanirost and Phumchusri, 2020).

However, high drop-off rates are not uncommon and have been studied previously. Walsh et al. (2020) investigated the occurrence of high drop-offs on a museum website. It was observed that high volumes of viewers would look at only one or two pages within 10 seconds of the visit, and thereafter, they would drop-off. The study probed a better understanding of the type of users who visited the website to explain the high drop-off rate. It was found that the majority of the drop-offs were linked to the understudied general public and non-professional users (Walsh et al., 2020).

Rojas et al. (2022) conducted a study to measure traffic and drop-offs in Colombian banking establishments. The researchers found that the websites of Colombian banking establishments were well positioned and presented low bounce and drop-off rates (Rojas et al., 2022).

Dou et al. (2018) claimed that users subconsciously assign a rapid and lasting impression on the attractiveness of the webpage within 50 milliseconds. Afterwards, the user's interest in and engagement with the website is highly

influenced by this subconscious assessment. To address and minimise website drop-offs, the study proposed a deep neural network model to compute and quantify the webpage aesthetics, which has proven to be an effective aesthetics evaluation tool during the web design process (Dou et al., 2018).

Within this chapter, the observed drop-offs were explored by solving a survival problem. The Cox proportional hazard and random survival forest models were employed. A key objective of the study sought to identify the hazards that have influenced the observed drop-off rate. The owners of the website invested resources in the development and maintenance of the website and thus found the high drop-off rate concerning. Once the drop-off hazards were identified, the owners of the website intended to address these hazards to ultimately increase online engagement. This study contributes to a unique method of understanding drop-offs. At the time of writing, no present literature existed on website drop-off hazards that were explored through a survival problem that could be found with the comparison between the Cox proportional hazard model and the random survival forest model. The study of website drop-offs is fairly new and of growing concern as the world becomes more digitally enabled (Awichanirost and Phumchusri, 2020). This chapter documents a detailed illustration of identifying website drop-off hazards that could be replicated by other corporates experiencing high volumes of website drop-offs.

Within this chapter, Section 6.2 details the methodologies employed within this study and discusses the underlying theoretical framework. Section 6.3 discusses the survival rates of the data and discusses the data censoring prior to the survival model construction. Section 6.4 discusses the survival models, and the results are discussed in Section 6.5.

6.2 Materials and methods

To investigate the concerning volume of web drop-offs, this study employed two popular survival analysis techniques. The Cox proportional hazard model and the random survival forest model were employed to determine the key hazards that drove web drop-offs on the studied website. The Cox proportional hazard model follows a regression algorithm, whilst the random sur-

vival forest model falls within the family of modern ensemble machine learning algorithms.

6.2.1 Drop-off literature

The event of people or objects dropping-off from given environments has been of great interest across several fields of study. This section discusses previous research on drop-offs from an educational program, sports participation, and musical participation. Gubbels et al. (2019) conducted a study to gain further insight into the risk factors associated with school absenteeism and permanent school drop-offs. The study synthesised 75 studies with 635 potential risk factors for school drop-offs. According to the study, factors such as a history of grade retention, low IQ, learning difficulties, and low academic achievements have been shown to hold high significance (Gubbels et al., 2019).

Eime et al. (2019) conducted a study to probe the factors behind people dropping-off from sports participation. The study utilised amalgamated data where participants were registered in one of eleven sporting associations. Participants were categorised based on demographics, and a comparison was conducted between registration volumes and participation volumes to better understand drop-offs. The study found that individuals playing multiple sports peaked between ages 5–14 and thereafter diminished as specialisation increased. However, the drop-off in community sports participation was a concern during adolescence, and policy recommendations have been made to address the concern (Eime et al., 2019).

Pitts and Robinson (2016) investigated the individuals dropping-off from classical musical participation. Through the use of interviews of current and past members to explore the themes of social acceptance, musical satisfaction, and personal confidence to establish how individual determination competed against the circumstances that would hinder musical activity. The study found that the role of music education was fundamental to lifelong participation. Furthermore, the study discussed the benefits of exposing all children to the experience and the understanding of making music (Pitts and Robinson, 2016).

6.2.2 Methodology flowchart

To illustrate the methodology followed within the study, Figure 6.1 provides a high-level flowchart of the data collection, processing, and modelling.

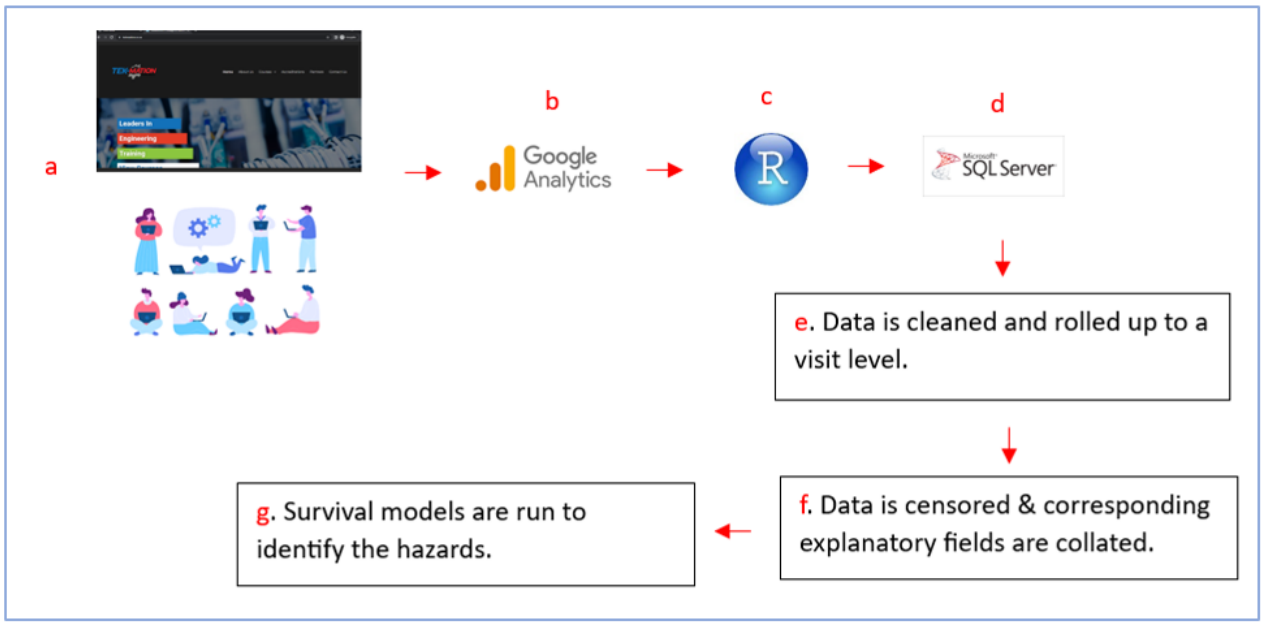

Figure 6.1: Web analytics data flow chart.

At a high-level, eight brief steps can be used to describe the methodology flow of the survival models:

i Visitors enter the studied website from their mobile devices, tablets, or computers.

ii The Google Analytics tracking tool tracks each user's activity and records important information (e.g., the geolocation of the device, the type of device or the detailed activity on the website).

iii Through an API, R (data science programming tool) was utilised to query the data from Google Analytics and shape the data into data frames.

iv The tracking data was then written out to a local SQL environment for further processing.

v Within SQL, the data was cleaned and rolled up to a visit level to summarise the entire engagement on the studied website per visit.

vi The data was censored to label the visits that were considered as survived.

vii Finally, the data then was fed into the model build and validation stages.

6.2.3 Kaplan-Meier curve

The Kaplan–Meier estimator represents a statistic (non-parametric) that is employed to estimate the survival function of survival (lifetime) data. The Kaplan–Meier estimator of the survival function (S_t), which represents the probability of survival (or life) being longer than time, t, is computed in Equation 6.1:

$$\hat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i}), \tag{6.1}$$

where t_i represents a point in time where at least one event has occurred, d_i counts the number of events that occurred at this point in time (t_i), and n_i denotes the total number of individuals that have survived up to time point (t_i) where $i \in t$ (D'Arrigo et al., 2021).

6.2.4 Cox proportional hazard regression

The fundamental theory of the Cox proportion hazard model holds that if the proportional hazards assumption is true, then without consideration of the full hazard model, it may be possible to estimate the effect parameters (α_i) as per Equation 6.2. The Cox proportional hazard model holds a hazard function of the form where X_i = ($X_{i1}, \ldots, X_{ip}$) as realised values of the explanatory variables for the subject i $\{i \in 1, \ldots, p\}$,

$$\lambda(t \mid X_i) = \lambda_0(t) e^{\alpha_1 X_{i1} + \ldots + \alpha_p X_{ip} + \alpha_0}, \tag{6.2}$$

which yields the hazard function at time point, t, for the subject, i, with the set of explanatory variables, X_i, and intercept term, α_0 (Kvamme et al., 2019). McLernon et al. (2022) attempted to assess the performance and usefulness of predictive survival outcomes through the use of a Cox proportional hazard

model. Ultimately, the recommendations of the study proposed a set of performance measures that can be used to validate predictions from the survival analysis model (McLernon et al., 2022).

Thiruvengadam et al. (2021) employed a Cox proportional hazards model to identify the factors that influence the duration of hospitalisation in COVID-19 patients. The Cox proportional hazard model's covariates that caused longer hospitalisation were abnormalities in oxygen saturation, neutrophil-lymphocyte, levels of D-dimer, lactate dehydrogenase, and ferritin. Furthermore, the findings of the study also indicated that patients with more than two chronic diseases had a significantly longer hospital stay (Thiruvengadam et al., 2021).

Matsuo et al. (2018) designed a survival problem to predict the survival outcome of cervical cancer patients. The study assessed the performance of the Cox proportional hazard model in comparison to deep learning models. The study found that on the observed data, the deep learning model outperformed the Cox proportional hazard model. On the studied features, it was apparent that the deep learning models outperformed the Cox proportional hazard model on the non-linear hazards (Matsuo et al., 2018).

Proportional property

Suppose there exists a single covariate, x, and a corresponding single coefficient α_i. Thus, consider the effect of incrementing x by 1, as per Equation 6.3 (Cox, 1997):

$$\begin{aligned} \lambda(t \mid x+1) &= \lambda_0(t)e^{\alpha_1(x+1)} \\ &= \lambda_0(t)e^{\alpha_1 x+\alpha_1} \\ &= \left(\lambda_0(t)e^{\alpha_1 x}\right)e^{\alpha_1} \\ &= \lambda_0(t \mid x)e^{\alpha_1}. \end{aligned} \tag{6.3}$$

Therefore, as per Equation 6.3, increasing the covariate x by 1, scales the original hazard by e^{α_1}, and by making e^{α_1} the subject of the formula in Equation 6.3 we obtain Equation 6.4:

$$\frac{\lambda(t \mid x+1)}{\lambda(t \mid x)} = e^{\alpha_1}. \tag{6.4}$$

From Equation 6.4, the right-hand-side is independent of time t. Since the incremental relationship of x is of the form $\frac{x}{y} = constant$, this is termed a proportional relationship.

Cox proportional hazard intercept

Unlike traditional regression models, the Cox proportional hazard regression model lacks an intercept term since the baseline hazard ($\lambda_0(t)$) accounts for the constant. Taking an intercept (α_0) (Cox, 1997):

$$\begin{aligned}
\lambda(t \mid x) &= \lambda_0(t) e^{\alpha_1 X_{i1} + \ldots + \alpha_p X_{ip} + \alpha_0} \\
&= \lambda_0(t) e^{X_i \cdot \alpha} e^{\alpha_0} \\
&= (e^{\alpha_0} \lambda_0(t)) e^{X_i \cdot \alpha} \\
&= \lambda_0^*(t) e^{X_i \cdot \alpha},
\end{aligned} \tag{6.5}$$

where $e^{\alpha_0}\lambda_0(t)$ in Equation 6.5 represents the revised baseline hazard $\lambda_0^*(t)$.

Likelihood of unique times

The likelihood (L_i) of an event to have been observed for a data point i at time T_i can be expressed as per Equation 6.6 (Cox, 1997):

$$\begin{aligned}
L_i(\alpha) &= \frac{\lambda(T_i|X_i)}{\sum_{j:T_j \geq T_i} \lambda(T_i|X_j)}, \\
&= \frac{\lambda_0(T_i)\theta_i}{\sum_{j:T_j \geq T_i} \lambda_0(T_i)\theta_j}, \\
&= \frac{\theta_i}{\sum_{j:T_j \geq T_i} \theta_j},
\end{aligned} \tag{6.6}$$

where $\theta_j = e^{X_j \cdot \alpha}$ and the sum over the set of data points j where the event has not occurred prior to time T_i with $0 < L_i\alpha \leq 1$. Therefore, Equation 6.6 represents the partial likelihood whereby the effects of the covariates could

be estimated with no need to necessarily model the variations of the hazard over time.

6.2.5 Random survival forest

The random survival forest model is a machine learning method that is employed to solve survival (time-to-event) problems (Wongvibulsin et al., 2019).

The algorithm follows the steps below:

1. The training dataset is boot-strapped into n subsets.
2. For each of the n sub-samples, a survival tree is composed where each node is randomly selected with $m \leq p$ where m represents the candidate number of variables considered and p is the total number of predictors. Of the m variables selected, the model would determine the optimal splitting of the variables and split points.
3. The model loops to continue recursive partitioning conditioned on no less than $d_0 > 0$ unique deaths.
4. Finally, compute the hazard function for the terminal nodes of the trees and determine the ensemble cumulative hazard function through a process of aggregation across the trees.

Jin et al. (2020) explored the use of random forests in survival analysis to understand employee attrition. The study proposed a hybrid model based on survival analysis and machine learning which combined survival analysis for censored data processing and training on attrition data patterns. The results of the study proved that the survival analysis model can materially improve the attrition prediction accuracy (Jin et al., 2020).

Soltaninejad et al. (2018) employed random forests to predict patient survival through the segmentation of brain tumours in multi-modal MRI images. Within the study, the classification accuracy, pairwise mean square error, and Spearman rank metrics were all acceptable (Soltaninejad et al., 2018).

Wongvibulsin et al. (2019) conducted a study to assess the prediction of clinical risk for survival using random forest models. The findings of the model highlighted the importance of features, such as the number of preceding heart failure hospitalisations (Wongvibulsin et al., 2019).

The Cox proportional hazard regression model, similar to linear regression and logistic regression, is linear in nature. Specifically, linear methods assume that a single line, curve, plane, or surface suffices to separate survival groups (alive, dead) or to estimate a quantitative response (survival time). However, alternative partitions may yield more accurate classification or quantitative estimates. A potential set of alternative methods are tree-structured survival models, including random survival forests (Ishwaran et al., 2008).

Bagging in random forests

The random survival forests training process follows the general routine of bootstrap aggregation (also known as bagging) to tree learning. Assume a training set $X = x_1, ..., x_n$ with responses $Y = y_1, ..., y_n$, bagging repeatedly (say B times) would extract a random sample with replacement of the training set and thereafter fits the trees onto these samples as per the steps below (Tinguely et al., 2019):

1. extract a sample (X_b, Y_b) of n (with replacement) from X, Y for training.
2. Thereafter, train a classification tree f_b on the sample (X_b, Y_b).

Upon completion of the training process, predictions are made on the unseen samples $\overline{x}$ aggregating the predictions from all the individual regression trees on $\overline{x}$:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(\overline{x}). \tag{6.7}$$

The process of bootstrapping yields higher model performance due to the minimisation of model variance, without increasing the bias. Therefore, although the predictions of a single tree are highly sensitive to any noise within the training subset, the average across many trees is far less influenced by noise. Training several trees on a single training dataset, would yield highly correlated trees, thus the need for bootstrapping to de-correlate the trees (but each tree training on a different set of data).

The uncertainty of the prediction can be estimated through the standard deviation of the predictions across all the individual regression trees on $\overline{x}$ (the test dataset):

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(\overline{x}) - \hat{f})^2}{B-1}}, \tag{6.8}$$

where the count of samples/trees, B, represents an unbound parameter. The optimal number of trees B could be determined through the cross-validation process, or by inspection of the out-of-bag error. The out-of-bag error represents the mean prediction error on each training sample x_i by selecting solely the trees that did not have x_i fall within the corresponding bootstrap sample.

Furthermore, random forests also include another type of bagging scheme on the features to split. Random forests employ a modified tree learning algorithm that chooses, at each candidate split in the learning process of the tree, a random subset of the features. This process is often termed feature bagging. The process of feature bagging ensures sufficient feature variation across the several trees. For instance, it may be possible that, if a few features are very strong predictors for the dependent variable, these features will be selected across most of the B trees, causing them to become correlated. As a rule of thumb, for a classification problem with p features, $\sqrt{p}$ features are used in each split. In regression random forest problems, the recommended rule of thumb is $\frac{p}{3}$ with a minimum node size of 5 as the default. However, in practice, the real-life best values for hyper-parameter tuning should be determined on a case-to-case basis for every problem (Hastie et al., 2009).

Variable importance

Random forest models are often employed to determine the features' importance in regression and classification applications. Two common techniques to determine the feature importance are the permutation importance and mean decrease in impurity feature importance.

The permutation importance is computed by firstly fitting a random forest to the dataset $D_n = \{(X_i, Y_i)\}_{i=1}^{n}$ where the dataset has $i \in (1, n)$ observations. The out-of-bag error is recorded and averaged over the random forest during

the fitting process. Thereafter, the feature importance of the $j-th$ feature after training is estimated through the values of permuted out-of-bag samples and the out-of-bag error is again computed on the perturbed dataset. Therefore, the importance score of the $j-th$ feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. Finally, the scores are then normalised.

The mean decrease in impurity feature importance considers features that decrease the impurity most during the splits, as important (Ortiz-Posadas, 2020).

$$average\ importance(x) = \frac{1}{n_T}\sum_{i=1}^{n_T}\sum_{node\ J \in T_i | split\ variable(j)=x} p_{T_i}(j)\Delta i_{T_i}(j) \quad (6.9)$$

where x represents the feature, n_T represents the number of trees within the forest, T_i represents the tree i, $p_{T_i}(j) = \frac{n_j}{n}$ represents the fraction of samples that reach the node j. The parameter $\Delta i_{T_i}(j)$ represents the change in impurity of the tree t at the node j.

6.2.6 Formal methods

During the deployment of predictive models, formal methods are often implemented to govern the predictive outcomes to minimise the impact of unexpected actions that could be the result of predictive errors (Krichen et al., 2022; Raman et al., 2023). Urban and Mine (2021) stressed the importance of software systems to behave correctly and reliably in safety-critical applications. For example, in avionics, the aircraft software has very stringent verification protocols that are mandatory as governed by international standards (Urban and Mine, 2021).

Within this study, the predictive survival model for website drop-offs had relatively low-risk consequences in the event of a model malfunction. However, to assess the accuracy of the Cox proportional hazard and the random survival forest models, the study randomly split the website visit data into a training subset (80%) and a testing subset (20%). The models were trained on the training set and thereafter validated on the test subset.

6.3 Exploratory analysis

This section discusses the exploratory analysis prior to the construction of the survival models. The data exclusions, definition of censoring, and Kaplan–Meier curves are discussed.

6.3.1 Webpage views

The studied data removed bounce visits from the survival analysis to better understand the drop-offs. By definition, a bounce visit would represent events of a person entering the website and thereafter, instantly dropping-off. This is often a symptom of a person mistakenly entering a website or very quickly determining upon landing on a website that it was not what the visitor was browsing for. In the survival problem, bounce visits would materially increase the model noise (Poulos et al., 2020). The owners of the studied website (a South African corporation), by design, intended that visitors should view a minimum of three pages per visit (after the exclusion of bounce visits). Figure 6.2 depicts the censored pageview distribution.

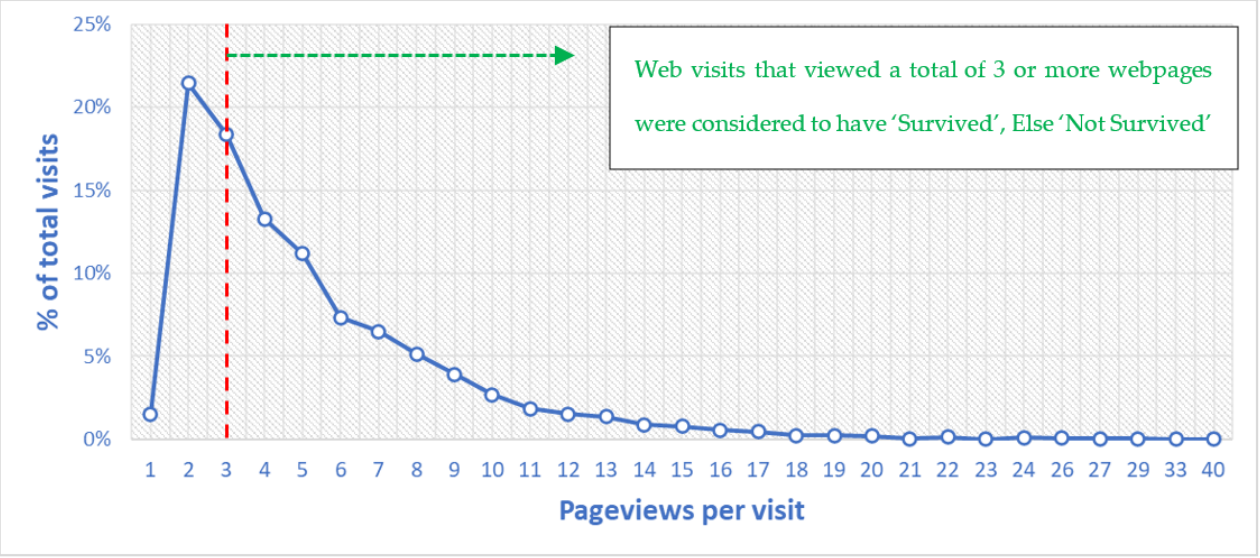

Figure 6.2: Survival analysis pageview distribution.

Each person entering the studied website, at a given instance, can view several webpages (e.g., *home*, *contact-us* or *about-us*). However, according to the observed data, roughly 22% of the visitors would enter the website and only

view two webpages within that website and thereafter drop-off. This implies that more than one in five visits did not survive to have viewed more than two webpages per visit. This study sought to identify the hazards that contribute to the material portion of visits that failed to survive through the use of a Cox proportional hazard model and a random survival forest model. Thereby, the definition of survival follows:

- visits that viewed in total three or more webpages, we considered survived.
- visits that viewed in total less than three webpages were considered a failure to survive.

6.3.2 Kaplan-Meier curve

A Kaplan–Meier curve was constructed to visually assess the survival rate as time progressed on the studied website (as depicted in Figure 6.3).

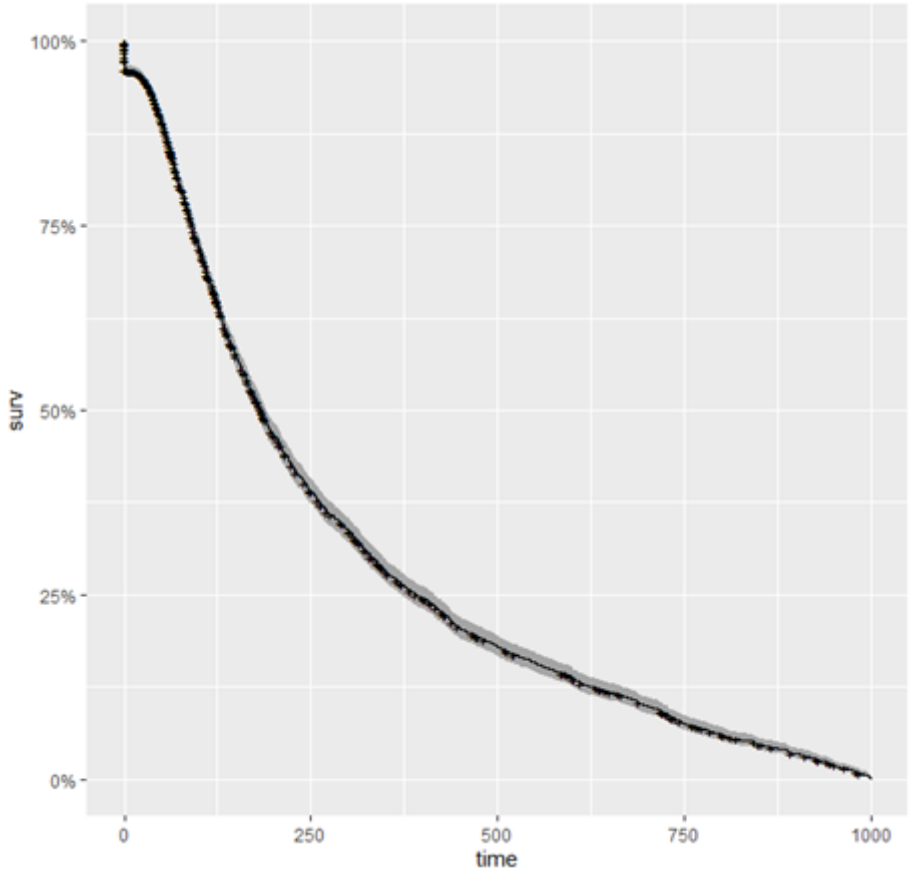

Figure 6.3: Survival analysis Kaplan-Meier curve.

On the studied website, the Kaplan–Meier survival curve followed a concave distribution which suggested that survival rates steeply decline within the first 250 seconds of a website visit. Additionally, after 250 seconds, a more gradual survival rate has been noted. This implies that the premature drop-offs leave the website shortly after entering, and those who view a greater number of pages have been more engaged with the website.

6.4 Survival models

Section 6.4 details the Cox proportional hazard model and random survival forest models. The features fed into the models are discussed, the features that the models have identified as hazards are discussed, and the models' classification accuracies are further shared.

6.4.1 Cox proportional hazard regression

A Cox proportional hazard model was employed to determine the significant hazards that contribute to survival on the studied website. The model sought to solve the problem as expressed in Equation 6.10 where survived referred to the event of a web visit viewing three or more webpages within the visit:

$$\begin{aligned} surv(sesstime, survived) \sim & browser + dayssincelastsession + \\ & devicecategory + distance + hits + \\ & operatingsystem + organicsearches + \\ & AC + AU + AP + CE + CL + CR + CU + \\ & EA + ET + H + L + SC + T + U. \end{aligned} \tag{6.10}$$

Prior to computing the Cox proportional hazard model, the categorical features (*operatingsystem*, *devicecategory*, and *browser*) were transformed into indicator variables. The H feature indicates if a visitor the *home* page of the website first. Similarly, the features *AC*, *AU*, *AP*, *CE*, *CL*, *CR*, *CU*, *EA*, *ET*, *L*, *SC*, *T* and *U* indicate the corresponding first webpage viewed within each visit as detailed in Table 2.2.

The Cox proportional hazard model was trained on 80% of the data and

tested on 20% of the data for validation. The Cox model recorded a classification accuracy of 58%. Figure 6.4 depicts the features and the corresponding hazard level of significance as determined by the Cox proportional model.

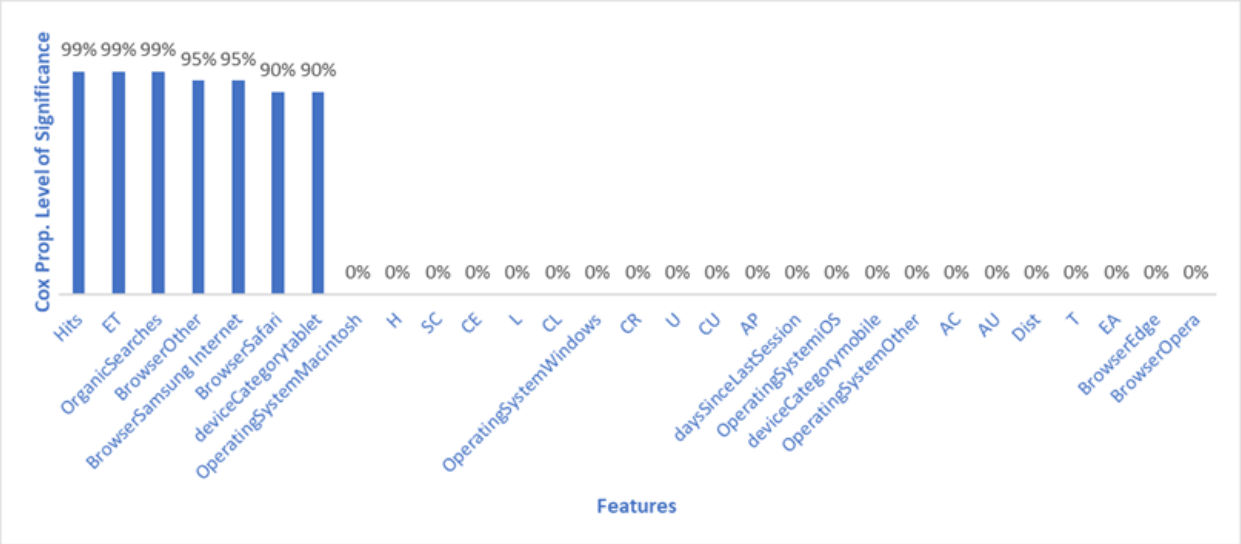

Figure 6.4: Web analytics survival Cox proportional hazards.

According to the Cox proportional hazard model, the features *hits*, *ET*, *organicsearches*, *browser*, and *devicecategory* were significant hazards to survival on the studied website. Further descriptive analysis showed that visits that were very engaged with the website (number of hits per page) had a high likelihood of survival. By definition, a hit represents the number of clicks a visitor makes when on a webpage. The feature *ET* represented visits where the first webpage viewed was the *engineering-trade* webpage. According to the data, visits with the first page being the *engineering-trade* have shown a high likelihood to drop-off early. This is suggestive that either such visits get all the information they need from the webpage or are dissatisfied and warranted to drop-off.

The data also informed that visits that originated from an organic search had a high likelihood of survival on the website, and conversely, those visits that were not organically originated showed a high likelihood to drop-off prior to viewing three or more webpages on the studied website. The Cox proportional model detected that visits from browsers such as Samsung Internet, Safari and other uncommon browsers had higher likelihoods of dropping-off prior to three page views. This could be due to either a population trait of

the users of such browsers (or devices associated with such web-browsers) or the visual presentation of the corporate website being non-functional on these browsers.

6.4.2 Random survival forest

The web traffic data has been modelled through a random survival forest model to identify the hazardous features that have contributed to visits not surviving three or more webpages on the studied website. The random forest's relative importance scores are depicted in Figure 6.5. The random survival forest model was trained on 80% of the data and tested on 20% of the data for validation. The random forest model recorded a classification accuracy of 63%.

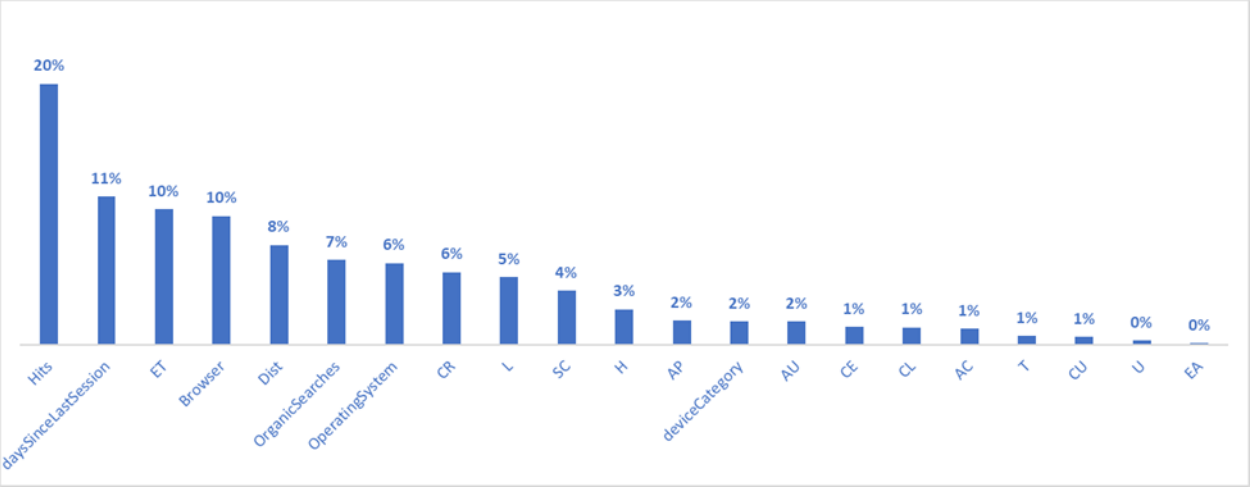

Figure 6.5: Web analytics random survival forest hazards.

According to the random survival forest model, *hits*, *dayssincethelastsession*, *ET*, *browser*, *distance* and *organicsearches* were among the most important features in determining if a visit would survive or not on the studied website. Empirical analysis has shown that the less engaging a visit was per webpage (measured by the volume of hits), the lower the chance of survival. The random forest model also detected that visits that were either very frequent or previously visited between 80–100 days ago have been shown to have a low survival rate. The survival model has also detected that the visitor's Euclidean distance between the device used to browse the website and the corporate's coordinates is an important hazard. The further the distance, the

lower the survival rate as expressed in the empirical data. This typically represented foreign visits. The *ET* feature has been identified as a hazard since visits with the *engineering-trade* page viewed first, have been shown to have a low survival rate on the studied data. The *browser* feature has been identified with high importance since visits with browsers, such as Samsung Internet, Safari and other uncommon browsers have been shown to have a low survival rate within the data. This could be an indication that either the owners of such devices have an unknown tendency to drop-off, or perhaps the visual display of the website on these browsers could contribute to a low survival. The *organicsearch* feature has also been identified as an important hazard, and according to the empirical data, it was evident that visits which have originated from an organic search have a high survival rate.

6.5 Final remarks

The shareholders of the studied website were concerned with the high rate of premature drop-offs. According to shareholder expectations, visitors should view three or more webpages per visit. A low volume of pageviews on the website implies that the marketing cost invested in the website is not being justified according to the intended purpose. This study employed a traditional survival model (through the use of the Cox proportional hazard model) and a machine learning model (random survival forest model) to identify the underlying hazards that drove the high rate of drop-offs observed. The Cox proportional hazard model was able to indicate features that were statistically significant in contributing to survival rates, whilst the random survival forest model was able to identify the features that are most important in predicting the survival rates. Although the models employed are fundamentally different, the features identified as hazardous were understandable and evident in the empirical data. Table 6.1 compares the features that were identified as hazardous under each model.

Table 6.1: Features identified as survival hazards.

Feature	Cox proportional hazard model	Random survival forest model
AC		
AP		
AU		
Browser	✓	✓
CE		
CL		
CR		
CU		
daysSinceLastSession		✓
deviceCategory	✓	
Dist		✓
EA		
ET	✓	✓
H		
Hits	✓	✓
L		
OperatingSystem		
OrganicSearches	✓	✓
SC		
T		
U		

Table 6.1 marks the features that were identified as possible hazards by the Cox proportional hazard model and the random survival forest model. Although both models are very different in theory, it was observed that a material overlap existed between the features identified as hazards. The features *browser*, *ET*, *hits* and *organic searches* were identified as hazardous by both the Cox proportional hazard model and the random survival forest model. The random survival forest model has been shown to better identify the hazardous features that were non-linear, such as *distance* and *dayssincethelastsession*. Furthermore, the Cox proportional hazard model identified the feature *devicecategory* whilst the random survival forest model did not.

The study illustrated how the hazards of a website could be successfully identified through a survival problem. The browser type (e.g., Google Chrome, Explorer or Firefox) was identified as a hazard and indicated that there could potentially be compatibility flaws with certain browsers that result in visi-

tors prematurely dropping-off. The developers of the website would need to re-test the website loading on the browsers of concern. The *engineering-trade* webpage was detected as hazardous and thereby, indicated that visitors entering the website through this webpage were perhaps discouraged by this webpage. Further tests would need to be conducted to isolate the underlying reason (through random A/B testing). Visitors could have either dropped-off due to the wording, due to the navigation to other webpages from this webpage, or due to the imagery, et cetera. The visitor's days since the last session were detected as a non-linear hazard. Exploratory analysis has shown that visitors who were new (first-time visitors) or visited a very long while ago have been shown to have had a higher survival rate. This implies that people who recently visited the website previously either entered again mistakenly or came on looking for specific information and thereafter dropped-off. The device category information was identified as a hazard as visitors using a desktop PC had a higher tendency to drop-off relative to tablet or mobile device users. Further research needs to be conducted to understand this behaviour. Further exploratory analysis has indicated that visitors from foreign countries enter the studied website and shortly dropped-off. The studied corporation at the time of analysis had only traded in and around South Africa and thereby, suggested that visits from outside of South Africa were most likely by mistake and thus explained the low survival rate as detected by a visitor's distance. The volume of hits has been identified as a hazard to survival as the degree of clicks and scrolls that a visitor has performed during the visit was a strong indicator of survival. Thus, intuitively, the lower the volume of hits, the lower the engagement, and provides an early indication of uninterest. Further testing should be performed to assess the impact of website prompts that can be thrown at a visitor with low engagement to try to increase the interest and engagement with the website to ultimately prevent premature drop-offs. Visitors who have not organically entered the website have proven to be hazardous. Organic searches represent visits whereby the visitor used keywords to describe what he was looking for, the search engine yielded its recommendations and thereafter the visitor chose the corporate website. Although this hazard may correlate with first-time visitors, further testing can be conducted to throw prompts to visitors who did not enter the website through an organic search to increase engagement and prolong the

web journey on the studied website. Ultimately, although both the Cox proportional hazard model and random survival forest models have performed well in identifying the key hazards, the random survival forest model has outperformed the Cox model due to its ability to better comprehend non-linear features and ultimately attained a superior classification accuracy relative to the Cox model. Although dropping-off has been a widely studied phenomenon across several applications, this study uniquely attempts to identify the underlying hazards by solving a survival problem. Given that the above methods have proved useful, and thus, web owners across several domains who experience high drop-off rates can likewise identify the underlying hazards by solving a survival problem.

Unlike Walsh et al. (2020) who studied the type of visitor that dropped-off, the use of a survival model can be generalised to any website and the outcome specifically indicates the hazards that need to be addressed. Dou et al. (2018) employed machine learning models to improve the visual appeal of the website to minimise drop-offs; however, visual appeal may not be the only factor that would result in premature drop-offs. For instance, the survival models have identified behavioural elements that may not be influenced by visual appeal alone (e.g., organically searched visitors or visitor dependency on the days since the last session).

Further to identifying the key hazards that drive premature drop-offs, a Markov chain model would highlight behavioural tendencies to dropping-off based on a visitor's present state within the visit journey. Chapter 7 employs tiered Markov chain models to quantify the likelihood of dropping-off based on the present state within a journey.

CHAPTER 7

PREDICT WEBPAGE TRANSITIONING USING MARKOV CHAINS

7.1 Introduction

In an attempt to model the most likely points of a web visitor prematurely dropping-off, Chapter 7 seeks to determine the probability of a visitor moving from one page to another. A website is made up of webpages where a visitor entering the website would navigate across several webpages by clicking on web objects. For example, a person would first enter the *home* page and thereafter click an object to transition to the *contact-us* webpage and thereafter exit the website. The underlying transition probability matrix would allow for practical implementation onto the live website. Thereby, whilst a visitor is on the website, the system could easily predict the most likely next state, and push pop-ups to walk a visitor through a desired page path.

Markov chains are based on the fundamental assumption that the future state is only dependent on the present state and not previous states (memorylessness). However, the memoryless assumption does not hold true on the studied website. This is intuitively so, as a visitor may not re-visit a webpage that has already been viewed. Therefore, to improve the accuracy of the Markov transition probabilities in the context of the studied website, a tiered Markov chain model was proposed.

The paragraphs below discuss recent applications of web behaviour prediction and Markov chain models. Koehn et al. (2020) have conducted a study to predict online shopping behaviour from clickstream information using deep learning methods. Their study found that a recurrent neural network and

conventional classifier models have captured the patterns inherent within the clickstream data. However, the study showed that ensemble methods consistently outperformed the alternate models tested (Koehn et al., 2020).

Nagaraj et al. (2023) employed machine learning models to predict e-commerce customer churn. The reported average monthly churn on the studied data was 2.2% and thus churn was not an easy event to predict (Nagaraj et al., 2023).

Rahman et al. (2019) employed a neuro-fuzzy approach to predict online behaviour using people's browsing interests and observing suspicious activities (e.g., security and privacy) derived from their internet trail. The proposed model was found to be promising in terms of the classification and prediction accuracy (Rahman et al., 2019).

Jia et al. (2020) employed Markov chain models to forecast coal consumption in the Gansu province. The final model was used to forecast coal consumption between 2020 and 2035 (Jia et al., 2020).

Vermeer and Trilling (2020) employed Markov chains using clickstream data of 175 news websites to better understand visitors. The outcome of their study proposed sales design strategies and guided on a more effective website structure (Vermeer and Trilling, 2020).

Okwuashi and Ndehedehe (2020) employed an integration between machine learning models and Markov chains to model the change in urban land usage. The final outcome resulted in a high accuracy level and proved to be a robust method for modelling urban change (Okwuashi and Ndehedehe, 2020).

Given the recent literature reviewed herein, none of the applications attempted to address cases where the memoryless property of the Markov model may not realistically apply. At the time of writing, no similar literature has been found that proposed a tiered Markov model in the manner as done within this study. The predictive model is required to be simplistic to allow easy implementation within the underlying code behind the website to ultimately lure visitors onto the more important webpages according to the TEKmation board. Although Markov models were considered to be simple enough (due

to the transition probability matrix), the memoryless assumption of Markov models was not ideal. The study aimed to evaluate a tiered approach to employing Markov models to minimise reliance on the memoryless assumption.

Section 7.2 of this Chapter discusses the theoretical framework of Markov chains, Section 7.4 discusses the Markov states data processing, Section 7.4 explains the derivation of the tiers and presents the tiered Markov models. Section 7.5 discusses the results of the models.

7.2 Markov methodology

This study investigated the use of tiered Markov chain models to minimise the assumption of memoryless transitioning. This section introduces the underlying mathematical theory behind discrete Markov chains and introduces related concepts such as absorbing states and time-homogeneity.

7.2.1 Discrete Markov chains

By definition, a discrete Markov chain represents a sequence of random variables $(X_1, X_2, X_3, \dots)$ that follow a Markov property (memoryless property), where the probability of moving to the next state $(n+1)$ depends only on the present state (n) and not on previous states:

$$P(X_{n+1} = x \mid X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(X_{n+1} = x \mid X_n = x_n) \quad (7.1)$$

where $P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) > 0$.

Thus, the possible values for X_i form a finite state space (T) of the chain (Cortes et al., 2020; Odhiambo et al., 2020).

7.2.2 Transition matrix

Assume a state space $(T = 1, \dots, d)$ which represents all possible states that a variable could reside in. The transition probability matrix would represent the probability of moving from one state (m) to the next (n) in one step. Therefore, $P(n \mid m) = T^{mn}$ could be described as (Wilinski, 2019):

$$\begin{pmatrix} T^{00} & \dots & T^{0d} \\ \dots & \dots & \dots \\ T^{d0} & \dots & T^{dd} \end{pmatrix}. \tag{7.2}$$

The transition probability distribution can be defined as, for any event $n \in (0, 1, 2, ...)$ and corresponding time values of n: $t_0, t_1, t_2, ...$ and respective states $i_0, i_1, i_2, i_3, ...$:

$$P(X_{t_{n+1}} = i_{n+1} \mid X_{t_0} = i_0, X_{t_1} = i_1, ..., X_{t_n} = i_n) = p_{i_n i_{n+1}}(t_{n+1} - t_n), \tag{7.3}$$

where p_{ij} represent the solution of the forward equation ($P'(t) = P(t)Q$) with initial condition $P(0)$ is the identity matrix where Q represents the transition matrix within the state space T.

7.2.3 Absorbing state

An absorbing state refers to a state that can be reached from any other state. Once in an absorbing state, the random variable cannot leave that state (Li and Ji, 2020).

In the context of this study, where a state refers to the webpage being viewed, at any given state a visitor could choose to subsequently leave the website (drop-off). However, once a person has dropped-off, the visitor cannot transition to any other state afterwards. After dropping-off, should the visitor re-enter the website at a later point, it would be treated as another visit.

7.2.4 Variations of Markov chains

Time-homogenous Markov chains

Time-homogenous Markov chains follow the assumption that transition probabilities do not depend on time 't'. Therefore, a Markov chain is said to be time-homogenous if (Mahmoudi and Rigi, 2023):

$$P(X_{t+1} = a \mid X_t = b) = P(X_1 = a \mid X_0 = b). \tag{7.4}$$

Stationary Markov chains

A stationary Markov chain represents a process where (Wilinski, 2019) :

$$P(X_0 = x_0, X_1 = x_1, ..., X_k = x_k) = \\ P(X_n = x_0, X_{n+1} = x_1, ..., X_{n+k} = x_k) \forall n, k. \tag{7.5}$$

A mandatory and sufficient criterion for time-homogeneous Markov chains to be deemed stationary requires the distribution of X_0 to be a stationary distribution of the Markov chain. Furthermore, the Baye's rule could prove that every stationary Markov chain is time-homogeneous.

Markov chains with memory (joint Markov chains)

By definition, a Markov chain of order m (or a Markov chain with memory), satisfies the following process (where m is finite):

$$P(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2,...,X_1=x_1}) \\ P(X_n = x_n \mid X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2,...,X_{n-m}=x_{n-m}}), \tag{7.6}$$

for n > m. Accordingly, as per Equation 7.6, the future states within a Markov chain with memory depends on the past m states. Therefore, it is possible to build a Markov chain (Y_n) from (X_n) which holds the traditional Markov property by defining the state space as the $m - tuples$ of X values: $Y_n = (X_n, X_{n-1}, ..., X_{n-m+1})$ (Wilinski, 2019).

7.3 Markov states

The tracking tool supplied data on the volume of visits to the website and the corresponding engagement whilst on the website. In the context of this study, the tracking tool recorded the webpages that a visitor had browsed on the studied website. Figure 7.1 below depicts a typical visit decision tree.

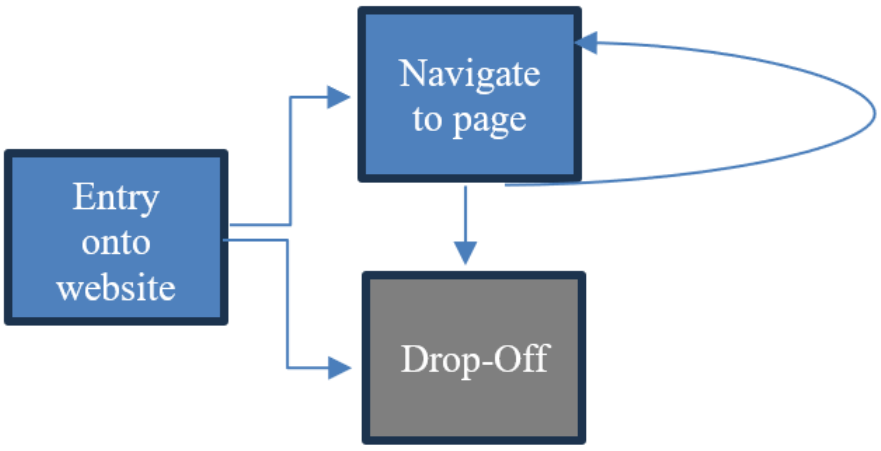

Figure 7.1: Online visit page decision tree.

The *entry-onto-website* block of Figure 7.1, represents the phase a person would enter the website on a given webpage. A person can enter the website on any given webpage (e.g., some may enter and land on the *home* page first, while others may land on the *contact-us* page first). Therefore, the order of pages viewed would vary across the visitors. Although most links would route to the *home* page first, web browsers (e.g., Google) typically yield a few options to the person prior to entering the website allowing visitors to land on specific pages of the website (e.g., directly to the *about-us* page). Furthermore, if a person has visited the website before, their last viewed page may be the entry point onto the studied website. The visitor would then decide to either drop-off, resulting in the visit ending, or view another page by clicking on relevant links. Subsequently, the *navigate-to-page* box in Figure 7.1 would loop until the person decides to exit.

The studied website was composed of several underlying webpages. To allow for a practical Markov chain application, the detailed pages were rolled up into the corresponding root pages. For example, all of the detailed courses pages that elaborated on educational courses offered by the corporation have been rolled up to the state: *courses*. The root pages can be viewed in Figure 7.2 for context.

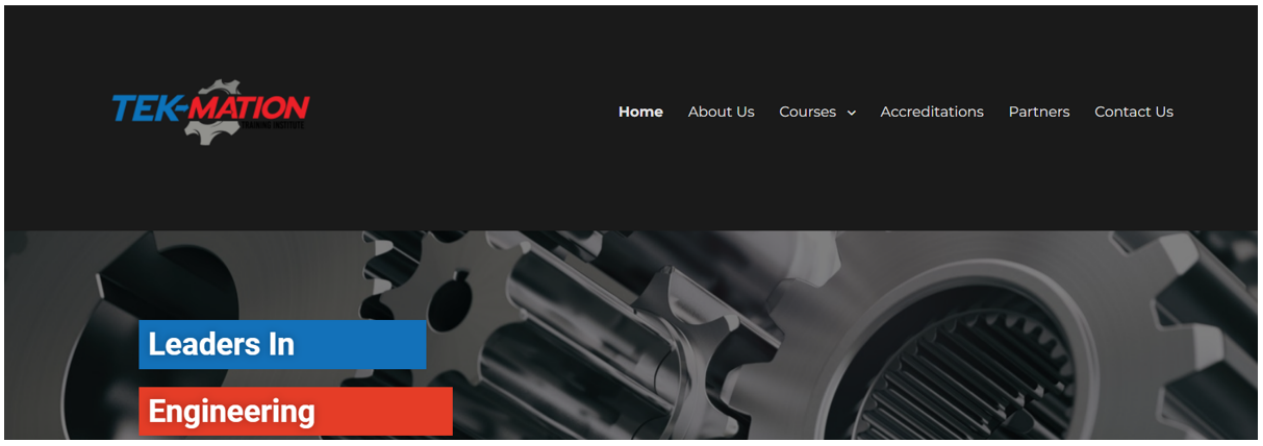

Figure 7.2: Studied website extract (source: Author's own contribution).

This way, within a visit, a person would reside in one of the six states and would subsequently either transition to one of the six states or leave the website (drop-off). The seven states modelled within this study are *about*, *accred*, *clients*, *contact*, *courses*, *home* and *drop-off* states.

7.4 Markov results

This section, firstly, tests the assumption of memoryless webpage transitioning through the use of chi-squared tests for independence. Thereafter, this section presents the distribution of webpages viewed to determine the number of tiers that were necessary. Finally, this section presents the transition matrices and Markov chains of website visits that transition from one state to the next through the use of tiered Markov models. The chi-squared test for independence and Markov chains were developed using the R data-science programming tool (package: markovchain (Spedicato et al., 2015)).

7.4.1 Webpage state historic dependence

Markov chain models are often termed memoryless models. This implies that the future state is dependent on the present state only and not the previous state. In the context of this study, the memoryless assumption would imply that the movement from one webpage to the next is dependent on the current webpage being viewed only and not dependent on the previous webpages view. However, this assumption does not hold true in the context of a

website as:

i. visitors are less likely to visit a webpage that they have viewed before relative to an unseen page, and

ii. the transition probability of the first webpage viewed would differ from the n^{th} – the first page may hold a higher drop-off rate as visitors may realise upon entry that the website was not what they were browsing for.

The chi-squared test for independence (with p-value < 0.0001) indicated that on the studied website, on a given webpage (w_0) the transition to the next webpage (w_1) is dependent on the previous webpage (w_{-1}). Since the p-value was less than the significance level of 0.05, we reject the null hypothesis and conclude that the next state was dependent on the previous state. Therefore, the memoryless transition probabilities of a Markov model would potentially dilute the probabilities.

7.4.2 Distribution of webpages viewed

On the studied website, roughly one in five visitors would drop-off on or before viewing the second webpage. Figure 7.3 depicts the distribution of the total count of webpages viewed on the studied website.

Figure 7.3: Distribution of the number of pages viewed per visit.

Using the distribution of pageviews, the evidence suggested that at minimum, a two-tier Markov chain model was necessary. The first tier would hold the transition probabilities of people whilst on their first and second webpage. The second tier would hold the transition probabilities of the people who were on their third or more webpage. Although adding more tiers may enhance the prediction accuracy of the next state, the researcher strove to keep the process as simple as possible to allow for practical application.

7.4.3 Tiered Markov models

Whilst a viewer is on a given webpage, to predict the subsequent webpage viewed, Markov chain models were developed. However, empirical results have shown that the process of web browsing is not a memoryless process. The next state (next webpage viewed) was shown to be dependent on the previous state (previous webpage viewed) whilst on the current state (webpage currently being viewed). Therefore, a tiered approach was proposed where the first tier represents the transition probabilities of the next state given that the visitor has recently entered the website (viewed less than three pages at that point). And thereafter, the second tier represents the transition probabilities when the current state was the third or more webpage. Table 7.1 below quantifies the transitional probabilities of the tier one Markov model.

Table 7.1: Tier one transition probabilities.

Tier 1	About	Accred	Clients	Contact	Courses	Home	DropOff
About	3%	8%	5%	20%	23%	10%	32%
Accred	2%	3%	13%	7%	46%	4%	25%
Clients	6%	4%	1%	13%	17%	10%	50%
Contact	2%	0%	1%	5%	18%	11%	63%
Courses	1%	2%	1%	2%	60%	6%	28%
Home	7%	4%	1%	7%	40%	8%	34%
DropOff	0%	0%	0%	0%	0%	0%	100%

The rows in Table 7.1 represent the current state (current webpage viewed) given that the current visit has just started (the current state is the first or second webpage of the visit) and the columns represent the probability of

transitioning to the next state (viewing the next webpage). According to the tier one transition probabilities, a visitor has a 20% probability of navigating to the *contact* state from the *about* state after the first two pages are viewed.

Figure 7.4 depicts the tier one Markov chains as represented by the transition probabilities.

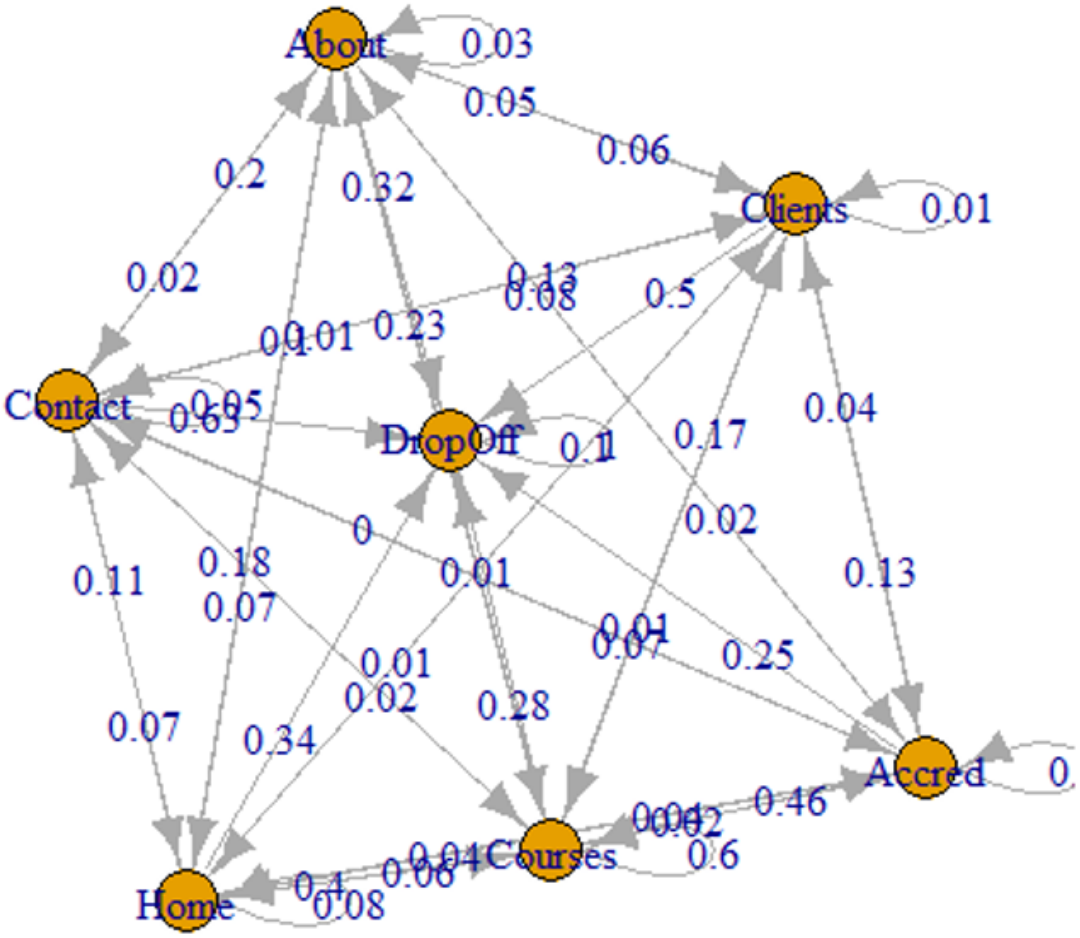

Figure 7.4: Tier one Markov chains.

The tier one Markov chains presented in Figure 7.4, depict the seven current states and quantifies the transition probability of moving to the next state. It can also be seen that the *drop-off* state is an absorbing state meaning that once the current state is at *drop-off* there will be no subsequent state. The *drop-off* state represents the probability of which a visitor would leave the website and thus would not be possible to transition to any other webpage. If the visitor re-enters the website, the event will be recorded as a separate visit.

Table 7.2 below presents the tier two transition probabilities on the studied website. These probabilities represent the likelihood that a visitor would transition to another state given the current page is the third or more web-page being viewed by the visitor.

Table 7.2: Tier two transition probabilities.

Tier two	About	Accred	Clients	Contact	Courses	Home	DropOff
About	3%	16%	7%	13%	36%	9%	16%
Accred	1%	2%	18%	10%	53%	5%	12%
Clients	4%	3%	1%	23%	40%	11%	18%
Contact	3%	1%	2%	4%	37%	13%	40%
Courses	2%	2%	1%	2%	61%	11%	21%
Home	11%	8%	3%	6%	48%	8%	17%
DropOff	0%	0%	0%	0%	0%	0%	100%

The rows of Table 7.2 represent the current state, and the columns represent the next state. According to the tier two transition probabilities, as per Table 7.2, there is a 13% chance of transitioning to the *contact* state from the *about* state. Figure 7.5 depicts the tier two Markov chains.

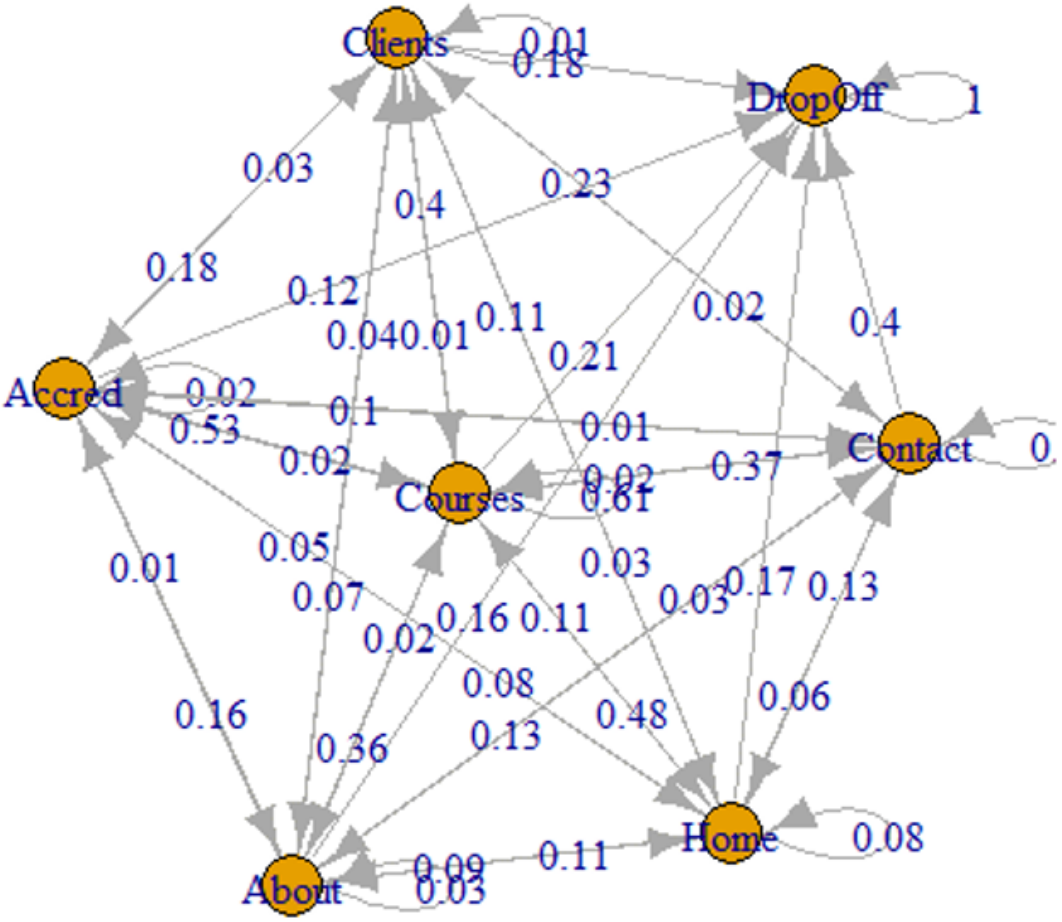

Figure 7.5: Tier two Markov chains.

Similar to the tier one Markov chain model, Figure 7.5 depicts the seven current states and quantifies the transition probability of moving to the next state.

7.4.4 Joint (hidden) Markov model

Since the transitioning from one page to the next does not follow a 'memoryless' process, a joint hidden Markov model of order two was developed to assess the prediction capability of transitioning from one state to the next. Table 7.3 presents the joint Markov chain transition probabilities when predicting the next state whilst on the current state given the previous state.

Table 7.3: Joint Markov chain model of order two.

Joint Markov model	About	Accred	Clients	Contact	Courses	Home	DropOff
About | Accred	0%	13%	6%	19%	19%	13%	31%
About | Clients	8%	16%	16%	4%	24%	16%	16%
About | Contact	2%	21%	7%	16%	23%	14%	16%
About | Courses	3%	17%	7%	11%	42%	10%	11%
About | Home	3%	11%	6%	18%	30%	10%	21%
About | New	5%	3%	3%	15%	13%	3%	58%
Accred | About	2%	4%	19%	13%	40%	10%	11%
Accred | Clients	5%	0%	0%	15%	55%	0%	25%
Accred | Contact	6%	6%	6%	19%	31%	19%	13%
Accred | Courses	1%	2%	18%	10%	55%	4%	11%
Accred | Home	2%	1%	17%	5%	55%	4%	16%
Accred | New	2%	4%	7%	7%	37%	4%	40%
Clients | About	3%	0%	0%	43%	28%	10%	16%
Clients | Accred	3%	3%	3%	27%	37%	10%	17%
Clients | Contact	9%	9%	0%	17%	22%	9%	35%
Clients | Courses	2%	3%	1%	12%	52%	11%	19%
Clients | Home	10%	3%	0%	21%	30%	14%	21%
Clients | New	0%	5%	2%	0%	7%	7%	80%
Contact | About	6%	1%	3%	5%	24%	13%	48%
Contact | Accred	6%	3%	2%	3%	49%	8%	28%
Contact | Clients	3%	2%	2%	4%	35%	15%	38%
Contact | Courses	2%	1%	2%	4%	43%	11%	36%
Contact | Home	2%	1%	1%	4%	25%	17%	51%
Contact | New	1%	0%	1%	6%	13%	5%	74%
Courses | About	3%	4%	2%	6%	54%	10%	20%
Courses | Accred	2%	2%	4%	3%	61%	10%	18%
Courses | Clients	3%	3%	2%	4%	51%	14%	23%
Courses | Contact	2%	2%	1%	3%	50%	11%	31%
Courses | Home	1%	2%	1%	2%	69%	9%	17%
Courses | New	1%	1%	0%	1%	41%	3%	53%
Home | About	22%	8%	4%	6%	27%	8%	26%
Home | Accred	17%	25%	4%	6%	35%	4%	8%
Home | Clients	10%	12%	7%	7%	48%	1%	13%
Home | Contact	13%	6%	3%	13%	28%	5%	32%
Home | Courses	10%	8%	3%	5%	56%	4%	14%
Home | New	6%	3%	1%	7%	40%	7%	35%

However, the joint hidden Markov chain model of order two yielded inferior results relative to the tiered models when predicting the transition to the

drop-off state. This was driven by the joint Markov chain's inability to determine the point of the visit journey when on the current state (unlike the tiered models proposed). For instance, the joint state of *Home* | *About* would represent visits where the *About* was the first state of the journey and when the *About* state could have been the tenth. This is due to a non-fixed order transition process of the web state transitions.

7.5 Concluding thoughts

The study sought to propose a tiered approach to a non-memoryless application of Markov models. With web analytics, Markov chain models and the underlying transition probabilities can be easily implemented within web-based applications to guide a visitor's journey on a given website and potentially minimise drop-offs. However, a major hurdle is the accuracy of the transition probabilities since web behaviour has proven to have memory. In other words, predicting the next page a visitor would select was indeed dependent on the previous pages that the visitor has already seen within a particular visit.

The general concerns of Markov chain models were addressed as tabulated in Table 7.4.

Table 7.4: General Markov chain model concerns.

Markov chain concern	Studies actions to mitigate the concern
Memorylessness assumption	It was established that the likelihood of transitioning from one webpage to the next was indeed dependent on the previous pages viewed and thus violating the memoryless assumption (Vermeer and Trilling, 2020). The memoryless assumption has been addressed through tiered Markov models which proved accurate. Given that most visits contained less than three webpage views (as per Figure 7.3), the tiers were defined by the first two pages viewed as tier one and the third or successive pages viewed under tier two. Upon comparing the drop-off probabilities across the transition matrices, the tiered models have shown to minimize the effect of the memorylessness assumption (Table 7.5).
Fixed order transition	A visit to the website could be initiated on a range of webpages (Kantanantha and Awichanirost, 2022; Kumar and Ogunmola, 2019; Onder and Berbekova, 2021). For instance, a visitor may begin his journey on the website by starting on the *home* page whilst another may begin on the *contact-us* page. And accordingly, the second, third and successive pages viewed would be different across visitors and therefore violating a fixed order transitioning process (as illustrated in Figure 7.1). To minimize the effect of a non-fixed order process, the tiered models separated early stages of a visit from later stages where the difference between the tiers were the orders. The performance of the tiers are presented within Table 7.1 and Table 7.2.
Sparse data and state explosion	The studied website contains around twenty unique webpages. Therefore, to ensure sufficient data and a controlled state set within the transition matrices (Kantanantha and Awichanirost, 2022; Kumar and Ogunmola, 2019; Onder and Berbekova, 2021), each webpage was mapped to its root webpage. In doing so, this resulted in six states (exclusive of a drop-off state). The definition of the six states is discussed within Section and illustrated in Figure 7.2. Therefore, sparse data and state explosion was managed by rolling up the web page visits into six states.

Upon studying the tier one and tier two transition probabilities, it was evident that the probability distributions were more discriminant. Tier one

highlighted that when visitors were on the *clients* or *contact-us* states early within their journey, they held a high likelihood of dropping-off. Table 7.5 quantifies the average probability of moving to the respective states between the tiered and non-tiered models.

Table 7.5: Average transition probability of the Markov models.

Avg. Prob	About	Accred	Clients	Contact	Courses	Home	DropOff
Non-Tiered	3%	4%	5%	10%	41%	9%	28%
Tier1	3%	3%	4%	9%	34%	8%	39%
Tier2	4%	5%	5%	10%	46%	9%	21%

The non-tiered model represents the Markov model that fully assumes that the future state is independent of previous states. Although the data has proven that the dependence does exist, the average transition probabilities of the non-tiered model were included for validation purposes. It is evident in Table 7.5, that the tiered models have resulted in greater differentiation relative to the non-tiered model. This implied that by introducing tiers, the assumption of previous state independence was reduced. Furthermore, according to the average probabilities, as per Table 7.5, the Tier one model would more accurately predict drop-offs. Furthermore, from the transition probabilities, it is evident that the tier two model would more accurately predict movement across the states relative to the tier one model.

Although the transition probabilities presented within the study were specific to the studied website, the methodology could likewise be applied to other websites to predict the subsequent most likely action of a web visitor.

On the studied website, the Markov transition probabilities could be coded into the website infrastructure to first, predict visitor drop-offs and push prompts to prolong the visit to the website. Secondly, the board of directors claim that certain webpages are more important than others and could therefore attempt to redesign less-important webpages to lure visitors onto the more important webpages. These changes would first be tested through an A/B test and thereafter rolled out.

CHAPTER 8

DISCUSSION AND CONCLUSION

8.1 Introduction

Web analytics is of high importance within the competitive global marketplace, as this is necessary to optimise aesthetics and user experience on a given website (Mbete and Tanamal, 2020; Rita et al., 2019; Wang and Law, 2019). Like TEKmation, many companies invest in an official website to represent the company digitally but they lack insight into the website's usage, the intents behind the visits, and the existence of any concerns about the website (Bleier et al., 2018; Noorbehbahani et al., 2019). Thereby, this study aimed to employ statistical and machine learning methods on the complex website data to yield a clear understanding of the underlying visit behaviour and identify any evidence of potential concerns.

The exploratory analysis conducted in this study, recorded a web traffic of 7,935 visits, with 5,997 unique visitors during the analysis period; these key metrics were in-line with the traffic observed in the study conducted by Jansen et al. (2022). After the data-shift, it was observed that the studied website experienced satisfactory traffic flow, with a relatively low bounce rate. However, a significant portion of visits was identified as premature drop-offs and thus warranted further investigation. The extreme values observed on the visit duration, the number of pages viewed per visit, and the number of times the website was visited were found to be within reason, though (although identified as outliers by Grubb's test). The small volume of missing values was imputed by using Bayesian network models.

A shift in data patterns was experienced during the harsh levels of the COVID-19 lock-down and the study proposed an artificial neural network model to

detect the likely presence of a future shift in browsing patterns. This artificial neural network model indicated that volume-based measures (e.g., visits and pageviews) were important predictors of a data-shift occurring, being cognisant of the time of year (due to seasonal periods). Although the data-shift period was omitted from the study, the learnings from the data-shift were extracted in the form of the artificial neural network model.

Using the web-tracking variables available, a feature selection process was conducted prior to the running of unsupervised machine learning models. The feature selection process focused on the variability of features and the association between features across various data types to identify features that contained poor variability and redundant information. Features such as the volume of pageviews and hits were identified as sharing a high correlation with each other, features such as the days since the last session and the user type were identified as features with natural relationships, and features such as visits to the engineering-trade webpage were identified as low variability features.

The study employed three unsupervised machine learning techniques to cluster visits based on web behaviour to identify the prevalent themes that were observed on the studied data. Across the clustering models, the clusters that emerged clearly identified significant groupings of intents behind the website visits. Each of the unsupervised machine models belongs to a different family of clustering techniques, namely K-means, DBScan and hierarchical clustering. However, across all three models, similar intents emerged, labelled: *get-in-touch*, *accidentals/drop-offs*, *seekers* and *engrossed*, which has proven very insightful and quantified the general behaviour on the website. It was clear that between 40% to 50% of the visits were *accidentals/drop-offs*, and 25% to 40% were found to be *engrossed*. The unsupervised machine learning models indicated that a material portion of the visits were drop-offs, as the visitors left the website prematurely (prior to viewing more than three webpages). To further model the volume of drop-offs, the study employed survival models and Markov chains.

The Cox proportional hazard model and random survival forest model identified the hazards that contributed to the premature drop-offs on the studied website; both models performed well in identifying these key hazards.

However, the random survival forest model out-performed the traditional Cox proportional hazard model in identifying the non-linear hazards. The analysis indicated that certain webpages, as well as the geo-location, level of website engagement, type of device and manner of entry, were significant indicators of the premature drop-offs.

Furthermore, the study employed tiered Markov chain models to predict the most likely subsequent action of visitors given their current webpage and point within the visit journey. In doing so, the most likely points of drop-off were determined. The tiered transition probabilities proved to be more discriminant and indicated that visitors who have landed on the *clients* and *contact-us* states early within the visit journey expressed a high possibility of dropping-off.

8.2 Future directions

The statistical and machine learning analysis conducted within this study has shared a detailed insight into the visitors of the website. The recommendations for TEKmation are as follows:

- Although the directors believed that the volumes were satisfactory, from the exploratory analysis and aggregated traffic flows on the website, further marketing campaigns should be considered to enhance traffic flow on the corporate website. Naturally, a higher volume of traffic on the website would result in higher awareness and greater potential business growth.
- The analysis indicated that a high proportion of visits ended prematurely. To minimise premature drop-offs, it is recommended that TEKmation conducts further A/B design tests by using web pop-ups to appear when the visitor portrays the hazardous behaviour as identified by the random survival forest model.
- The engineering-trade webpage needs to be re-designed as both survival models have indicated that the webpage was harmful to the business.

- It is suggested that TEKmation conducts further A/B design tests by using web pop-ups on the tier-one visits of the *clients* and the *contact-us* states, which indicated high drop-off transition probabilities.
- The study has been able to cluster the prevailing intentions behind website visits. It is recommended that TEKmation begins a journey to increase the proportion of engrossed visits without decreasing the total volume of visits.

With the recommendations presented, it is important that the volume of total visits remain the same or grow simultaneously. For instance, in the attempt to reduce premature drop-offs, close monitoring will need to be conducted to ensure that all other visits are unharmed. The statistically recommended method of introducing such changes is to use an A/B test in which a random 50% of the visits are passed through to the current website structure (the control population) and the remaining 50% to the revised website (test population). This will enable side-by-side comparisons to assess whether the intended revisions are working as required without causing any additional harm. It is also recommended that one change be made at a time to eliminate interaction effects between several simultaneous changes. Once the revisions have proven to be successful, they can be rolled out to all the website visitors (Howard and Ramdas, 2022; Koning et al., 2022).

In additional, future statistical recommendations include

- The study employed Markov chain models to predict website drop-offs. Perhaps other ensemble methods such as random forest models could be attempted. Markov models were employed due to the simplicity of embedding the transition probabilities into the web design. However, the Markov models may not have yielded the best accuracy although employed due to the practicality.
- The performance of the models employed were satisfactory on the studied website. However, further application of such methods on a variety of websites is recommended.

8.3 Limitations of the study

Although the statistical and machine learning methods applied may be useful to several other corporates attempting to analyse web traffic data, the limitations of the study and future research are discussed below.

On the studied website, a key assumption was that the IP address (a unique identifier assigned to every device by the internet service provider) represented a ClientID which was assumed to be a single person. However, a single person could have visited the studied website across different devices (each with a different IP address) and in doing so, the visits would be recorded under different ClientIDs. Although this assumption can be accommodated, since the visit behaviour of a person may differ across devices, the difference in behaviour between devices of a single person could not be explored in this study. Given this limitation, it is recommended that such analysis be conducted in a study with a login facility, thereby allowing to analyse user-behaviour across devices.

There are concerns about the privacy of tracking tools that closely monitor a visitor's activity on a given website (Winklbauer and Horner, 2022). Although web tracking agents (e.g., Google Analytics) ensure visitors are warned that their behaviour will be tracked on a particular website (visitors are requested to accept the use of cookies), many people are uncomfortable with the idea. Thereby, the future of web analytics may be uncertain.

A Bayesian network model was employed to accurately impute the low degree of missing data observed within the study on the mobile device branding and operating system features. However, the accuracy of the Bayesian network imputation leveraged off the completeness of the visit across all other features; in other words, the Bayesian model was able to predict the mobile device branding accurately, given that the browser type and the type of device were known. Further research could explore the imputation of data points with several such features simultaneously missing on the same record that was not present in the observed data. Finally, it should be noted that the Bayesian models have performed well on the high-incidence but not the low-incidence data points.

In conclusion, this study and the corresponding findings were limited to one particular website, future work assessing the online behaviours across several websites spanning different industries is recommended. Although the techniques applied within this study could be employed elsewhere, such future work would be of great value in understanding general browsing habits.

REFERENCES

Abraham, A. (2005). *Artificial neural networks*. John Wiley and Sons, Ltd Chichester, UK.

Adams-Cohen, N. (2020). Policy change and public opinion: Measuring shifting political sentiment with social media data. *American Politics Research*, *48*, 612–621.

Al-Lami, G. (2021). E-commerce: advantages and limitations. *International Journal of Academic Research in Accounting Finance and Management Sciences*, *11*, 153–165.

Alhlou, F., Asif, S., and Fettman, E. (2016). *Google Tag Manager Concepts*, (pp. 91–123).

Ali Mohammadi, H., and Chen, S. (2021). Performance evaluation of outlier detection techniques in production timeseries: A systematic review and meta-analysis. *Expert Systems with Applications*, *191*, 116371.

Andre, Q. (2021). Outlier exclusion procedures must be blind to the researcher's hypothesis. *Journal of Experimental Psychology: General*, *151*.

Arora, S., Bahukhandi, k. D., and Mishra, P. K. (2020). Coronavirus lockdown helped the environment to bounce back. *Science of The Total Environment*, *742*, 140573.

Atilgan, Y., Bali, T., Demirtas, K. O., and Gunaydin, A. (2019). Left-tail momentum: Underreaction to bad news, costly arbitrage and equity returns. *Journal of Financial Economics*, *135*(3).

Awan, M., Shafry, M., Rahim, M., Nobanee, H., Yasin, A., Khalaf, I., Ishfaq, U., and Javed, M. (2021). A big data approach to black friday sales. *Intelligent Automation and Soft Computing*, *27*, 785–797.

Awichanirost, J., and Phumchusri, N. (2020). Analyzing the effects of sessions on unique visitors and unique page views with google analytics: A case study of a tourism website in thailand. (pp. 1014–1018).

Bagirov, A. M., Aliguliyev, R. M., and Sultanova, N. (2023). Finding compact and well-separated clusters: Clustering using silhouette coefficients. *Pattern Recognition*, *135*, 109144.

Baycan, T., and Tuysuz, S. (2022). Special feature on social, economic, and spatial impacts of covid-19 pandemic in turkey. *Asia-Pacific Journal of Regional Science*, *6*.

Becker, L., and Gould, E. (2019). Microsoft power bi: Extending excel to manipulate, analyze, and visualize diverse data. *Serials Review*, *45*, 1–5.

Belair-Gagnon, V., and Holton, A. (2018). Boundary work, interloper media, and analytics in newsrooms: an analysis of the roles of web analytics companies in news production. *Digital Journalism*, *6*, 1–17.

Belhaj, A., Elraies, K., Alnarabiji, M., Abdulkareem, F., Muhamad Shuhili, J. A., Mahmood, S., and Belhaj, H. (2021). Experimental investigation, binary modelling and artificial neural network prediction of surfactant adsorption for enhanced oil recovery application. *Chemical Engineering Journal*, *406*, 127081.

Ben-Gal, I. (2008). *Bayesian networks*, vol. 1. Wiley Online Library.

Bentameur, K., and Belmihoub, I. (2022). *Advances in Search Engine Optimization Through Web Analytics Development: GuinRank's Web Analytics Case Study*, (pp. 321–332).

Bhrammanachote, W., and Sawangdee, Y. (2021). Sustaining or surviving? an exploratory case study on covid-19's impact towards hotel businesses. *Tourism and Hospitality Management*, *27*, 273–292.

Bibartiu, O., Durr, F., Rothermel, K., Ottenwalder, B., and Grau, A. (2023). Availability analysis of redundant and replicated cloud services with bayesian networks. *Quality and Reliability Engineering International*, *40*(1), 561–584.

Bleier, A., Harmeling, C., and Palmatier, R. (2018). Creating effective online customer experiences. *Journal of Marketing*, *83*, 002224291880993.

Brown, C. E. (1998). *Coefficient of Variation*. Berlin, Heidelberg: Springer Berlin Heidelberg.

Buschjager, S., Honysz, P.-J., and Morik, K. (2022). Randomized outlier detection with trees. *International Journal of Data Science and Analytics*, *13*, 1–14.

Cao, L., Zhao, Z., and Wang, D. (2023). *Clustering Algorithms*, (pp. 97–122). Singapore: Springer Nature Singapore.
URL `https://doi.org/10.1007/978-981-99-1533-0_5`

Chattopadhyay, A., Lee, C.-Y., Shen, Y.-C., Lu, K.-C., Hsiao, T.-H., Lin, C.-H., La, L.-C., Tsai, M.-H., Lu, T.-P., and Chuang, E. (2023). Multi-ethnic imputation system (mi-system): A genotype imputation server for high-dimensional data. *Journal of Biomedical Informatics*, *143*, 104423.

Chormunge, S., and Jena, S. (2018). Correlation based feature selection with clustering for high dimensional data. *Journal of Electrical Systems and Information Technology*, *5*, 542–549.

Chowdhry, A. (2023). Multiple imputation of missing data in practice. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *186*.

Cintron, D. (2022). *Bayesian networks*, (pp. 119–129). Elsevier.

Cirlugea, M., Farago, P., and Hintea, S. (2020). Statistical study of small business customers using facebook ads and google analytics. In *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, (pp. 212–215).

Collart, F., and Guisan, A. (2023). Small to train, small to test: Dealing with low sample size in model evaluation. *Ecological Informatics*, *75*, 102106.

Cortes, J., El-Labany, S., Navarro-Quiles, A., Selim, M., and Slama, H. (2020). A comprehensive probabilistic analysis of approximate sir-type epidemiological models via full randomized discrete-time markov chain formulation with applications. *Mathematical Methods in the Applied Sciences*, *43*.

Cox, D. R. (1997). Some remarks on the analysis of survival data. In *Proceedings of the first Seattle Symposium in Biostatistics: Survival Analysis*, (pp. 1–9). Springer.

Crutcher, P. D., Singh, N. K., and Tiegs, P. (2021). *Operating System*, (pp. 81–131). Berkeley, CA: Apress.

D'Arrigo, G., Leonardis, D., Abd Elhafeez, S., Fusaro, M., Tripepi, G., and Roumeliotis, S. (2021). Methods to analyse time-to-event data: The kaplan-meier survival curve. *Oxidative Medicine and Cellular Longevity*, *2021*, 1–7.

de la Cruz, R., and Kreft, J.-U. (2018). Geometric mean extension for data sets with zeros. *ArXiv*, (p. 1).

Dockes, J., Varoquaux, G., and Poline, J.-B. (2021). Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience*, *10*, 1–17.

Domazet, I., and Simovic, V. (2020). The use of google analytics for measuring website performance of non-formal education institution. *Handbook of Research on Social and Organizational Dynamics in the Digital Era*, *1*(1), 483–498.

Dou, Q., Zheng, S., Sun, T., and Heng, P.-A. (2018). Webthetics: Quantifying webpage aesthetics with deep learning. *International Journal of Human-Computer Studies*, *124*.

Dy, G. D., and Bordley, C. (2004). Feature selection for unsupervised learning. *Journal of Machine Learning Research*, *5*, 845–889.

Eime, R., Harvey, J., and Charity, M. (2019). Sport drop-out during adolescence: is it real, or an artefact of sampling behaviour? *International Journal of Sport Policy and Politics*, *11*, 1–12.

Enders, C. (2023). Missing data: An update on the state of the art. *Psychological Methods*, *1*, 1–10.

Ferreira, A., and Figueiredo, M. (2012). Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters*, *33*, 1794–1804.

Fop, M., and Murphy, T. (2017). Variable selection methods for model-based clustering. *Statistics Surveys*, *12*, 18–65.

Fraiman, R., Justel, A., and Svarc, M. (2008). Selection of variables for cluster analysis and classification rules. *Journal of the American Statistical Association*, *103*, 1294–1303.

Fu, H., Manogaran, G., Wu, K., Cao, M., Jiang, S., and Yang, A. (2019). Intelligent decision-making of online shopping behavior based on internet of things. *International Journal of Information Management*, *50*.

Galhotra, B., and Dewan, A. (2020). Impact of covid-19 on digital platforms and change in e-commerce shopping trends. *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, (pp. 861–866).

Garg, M., and Dhari, V. (2019). A regression model for analysis of bounce rate using web analytics. *International Journal of Innovative Technology and Exploring Engineering*, *8*, 646–649.

Gomer, B., and Yuan, K.-H. (2022). *Missing data analysis*. Elsevier.

Gubbels, J., Put, C., and Assink, M. (2019). Risk factors for school absenteeism and dropout: A meta-analytic review. *Journal of Youth and Adolescence*, *48*.

Guerif, S. (2008). Unsupervised variable selection: when random rankings sound as irrelevancy. *Journal of Machine Learning Research - Proceedings Track*, *4*, 163–177.

Guo, F., Ng, M., Goubran, M., Petersen, S., Piechnik, S., Neubauer, S., and Wright, G. (2020). Improving cardiac mri convolutional neural network segmentation on small training datasets and dataset shift: A continuous kernel cut approach. *Medical Image Analysis*, *61*, 101636.

Haji, K. (2021). E-commerce development in rural and remote areas of brics countries. *Journal of Integrative Agriculture*, *20*, 979–997.

Hanafi, N., and Saadatfar, H. (2022). A fast dbscan algorithm for big data based on efficient density calculation. *Expert Systems with Applications*, *203*, 117501.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer.

Helu, M., Sprock, T., Hartenstine, D., Venketesh, R., and Sobel, W. (2020). Scalable data pipeline architecture to support the industrial internet of things. *CIRP Annals*, *69*.

Holt, D., Scott, A., and Ewings, P. (1980). Chi-squared tests with survey data. *Journal of the Royal Statistical Society. Series A (General)*, *143*, 303–320.

Hong, M., and Lu, Z. (2023). *Multilevel Reliabilities with Missing Data*. Springer.

Hopfield, J. (1988). Artificial neural networks. *Circuits and Devices Magazine, IEEE*, *4*, 3 – 10.

Howard, S. R., and Ramdas, A. (2022). Sequential estimation of quantiles with applications to a/b testing and best-arm identification. *Bernoulli*, *28*(3), 1704–1728.

Hu, J., and Haddud, A. (2020). *Exploring the Impact of Globalization and Technology on Supply Chain Management: A Case of International E-Commerce Business*, (pp. 1353–1376).

Huang, J., Zhu, H., Liu, M., Zhang, T., and Wang, J. (2022). Achieving fast page load for websites across multiple domains. *Transactions on Emerging Telecommunications Technologies*, *33*.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, *2*(3), 841 – 860.

Jameel, K., Yazdeen, A., Ibrahim, A., Abdulrazzaq, M., and Mahmood, M. (2022). Analyses the performance of data warehouse architecture types. *Applied Soft Computing*, (pp. 45–57).

Jansen, J., Jung, S.-G., and Salminen, J. (2022). Measuring user interactions with websites: A comparison of two industry standard analytics approaches using data of 86 websites. *PLOS ONE*, *17*, e0268212.

Jia, Z.-q., Zhou, Z.-f., Zhang, H.-j., Li, B., and Zhang, Y.-x. (2020). Forecast of coal consumption in gansu province based on grey-markov chain model. *Energy*, *199*, 117444.

Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., and Qiang, B. (2020). *RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis*. Springer, Cham.

Johnson, J., and Khoshgoftaar, T. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*, 27.

Jonathan, O., Misra, S., Ibanga, E., Maskeliunas, R., Damasevicius, R., and Ahuja, R. (2019). Design and implementation of a mobile webcast application with google analytics and cloud messaging functionality. *Journal of Physics: Conference Series*, *1235*, 012023.

Joshi, M. Y., Rodler, A., Musy, M., Guernouti, S., Cools, M., and Teller, J. (2022). Identifying urban morphological archetypes for microclimate studies using a clustering approach. *Building and Environment*, *224*, 109574.

Kabisa, M., Subakanya, M., Malambo, M., Chapoto, A., Maredia, M., and Tshirley, D. (2021). Impact of covid-19 on household incomes and food consumption - the zambian case. *Research in agricultural and applied economics*, *1*(1), 1–9.

Kader, G., and Perry, M. (2007). Variability for categorical variables. *Journal of Statistics Education*, *15*.

Kalogiannidis, S. (2020). Covid impact on small business. *International Journal of Social Science and Economics Invention*, *6*.

Kalyankar, V., and Anute, N. (2022). A study on the effectiveness of google analytics on the business growth of e-commerce companies in india. *Journal of Information Technology and Sciences*, *8*(3), 1.

Kantanantha, N., and Awichanirost, J. (2022). Analyzing and forecasting online tour bookings using google analytics metrics. *Journal of Revenue and Pricing Management*, *21*, 354–365.

Kavzoglu, T., and Mather, P. (2002). The role of feature selection in artifial neural network applications. *International Journal of Remote Sensing*, *23*, 2919–2937.

Khan, M., Zubair, S. S., and Malik, M. (2019). An assessment of e-service quality, e-satisfaction and e-loyalty: Case of online shopping in pakistan. *South Asian Journal of Business Studies*, *8*(3), 283–302.

Kiley, K., and Vaisey, S. (2020). Measuring stability and change in personal culture using panel data. *American Sociological Review*, *85*, 477–506.

Koehn, D., Lessmann, S., and Schaal, M. (2020). Predicting online shopping behaviour from clickstream data using deep learning. *Expert Systems with Applications*, *150*, 113342.

Koning, R., Hasan, S., and Chatterji, A. (2022). Experimentation and start-up performance: Evidence from a/b testing. *Management Science*, *68*(9), 6434–6453.

Krenker, A., Bešter, J., and Kos, A. (2011). Introduction to the artificial neural networks. *Artificial Neural Networks: Methodological Advances and Biomedical Applications. InTech*, (pp. 1–18).

Krichen, M., Mihoub, A., Alzahrani, M. Y., Adoni, W. Y. H., and Nahhal, T. (2022). Are formal methods applicable to machine learning and artificial intelligence? In *2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, (pp. 48–53).

Kularia, S., Kumar, V., and Duhan, D. (2023). Evaluation of imputation techniques for genotypic data of soybean crop under missing completely at random mechanism. *Indian Journal of Agricultural Research*.

Kumar, J. A., Osman, S., Sanmugam, M. A., and Rasappan, R. (2022). Mobile learning acceptance post pandemic: A behavioural shift among engineering undergraduates. *Sustainability*, *14*, 3197.

Kumar, V., and Ogunmola, G. (2019). Web analytics for knowledge creation: A systematic review of tools, techniques, and practices. *International Journal of Cyber Behavior, Psychology and Learning*, *10*, 1–14.

Kvamme, H., Borgan, O., and Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, *20*, 1–30.

Li, T., Rezaeipanah, A., and El Din, E. M. T. (2022). An ensemble agglomerative hierarchical clustering algorithm based on clusters clustering technique and the novel similarity measurement. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 3828–3842.

Li, Y., and Ji, W. (2020). Understanding the dynamics of information flow during disaster response using absorbing markov chains. In *2020 Winter Simulation Conference (WSC)*, (pp. 2526–2535).

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451–461.

Luo, Y. (2021). New oli advantages in digital globalization. *International Business Review*, *30*, 101797.

Mahmoudi, P., and Rigi, A. (2023). Probabilistic prediction of drought in iran using homogenous and nonhomogenous markov chains. *Journal of Hydrologic Engineering*, *28*.

Mahmutovic, K. (2020). From google analytics to digital marketing optimization in hotel industry: Proposal of framework and empirical evaluation of hotel industry in croatia, bosnia and herzegovina and serbia.

Mariyapillai, J., and Pratheepan, T. (2021). Application of google analytics model for evaluating the visibility of library web portals of the uva wellassa university, sri lanka. *Emperor International Journal of Library and Information Technology Research*, *1*.

Markoulidakis, J., Kopsiaftis, G., Rallis, I., and Georgoulas, I. (2021). Multiclass confusion matrix reduction method and its application on net promoter score classification problem. *Technologies*, (pp. 412–419).

Matsuo, K., Purushotham, S., Jiang, B., Mandelbaum, R., Takiuchi, T., Liu, Y., and Roman, L. (2018). Survival outcome prediction in cervical cancer: Cox models versus deep-learning model. *American Journal of Obstetrics and Gynecology*, *220*.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009). Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics and Data Analysis*, *53*, 3872–3882.

Mbete, G., and Tanamal, R. (2020). Effect of easiness, service quality, price, trust of quality of information, and brand image of consumer purchase decision on shopee online purchase. *Jurnal Informatika Universitas Pamulang*, *5*, 100.

McGuirk, M. (2023). Performing web analytics with google analytics 4: a platform review. *Journal of Marketing Analytics*, (pp. 1–15).

McLernon, D., Giardiello, D., Van Calster, B., Wynants, L., Geloven, N., van Smeden, M., Therneau, T., and Steyerberg, E. (2022). Assessing performance and clinical usefulness in prediction models with survival outcomes: Practical guidance for cox proportional hazards models. *Annals of Internal Medicine*, *176*.

Meier, A., and Kaufmann, M. (2019). *SQL & NoSQL Databases: Models, Languages, Consistency Options and Architectures for Big Data Management*. Springer.

Miao, M., Jalees, T., Zaman, S. I., Khan, S., Hanif, N.-u.-A., and Javed, M. K. (2022). The influence of e-customer satisfaction, e-trust and perceived value on consumer's repurchase intention in b2c e-commerce segment. *Asia Pacific Journal of Marketing and Logistics*, *34*(10), 2184–2206.

Mitchell, S., Lahiff, A., Cummings, N., Hollocombe, J., Boskamp, B., Field, R., Reddyhoff, D., Zarebski, K., Wilson, A., Viola, B., Burke, M., Archibald, B., Bessell, P., Blackwell, R., Boden, L., Brett, A., Brett, S., Dundas, R., Enright, J., and Reeve, R. (2022). Fair data pipeline: provenance-driven data management for traceable scientific workflows. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *380*.

Mohammed, M., Naji, T., and Abduljabbar, H. (2019). The effect of the activation functions on the classification accuracy of satellite image by artificial neural network. *Energy Procedia*, *157*, 164–170.

Moreno-Barea, F., Jerez, J., and Franco, L. (2020). Improving classification accuracy using data augmentation on small data sets. *Expert Systems with Applications*, *161*, 113696.

Mudahemuka, W., Matundura, M., Murorunkwere, J., Kabera, C., Mbanzabugabo, J., and Ingabire, J. (2023). The social and economic impact of the covid-19 pandemic on private higher education in rwanda. *Journal of Research Innovation and Implications in Education*, (pp. 326–336).

Mustafa, S., Hao, T., Qiao, Y., Kifayat Shah, S., and Sun, R. (2022). How a successful implementation and sustainable growth of e-commerce can

be achieved in developing countries; a pathway towards green economy. *Frontiers in Environmental Science*, *10*, 1.

Nagaraj, P., Muneeswaran, V., .A, D., Aakash, M., Balananthanan, K., and Rajkumar, C. (2023). E-commerce customer churn prediction scheme based on customer behaviour using machine learning. *2023 International Conference on Computer Communication and Informatics (ICCCI)*, (pp. 1–6).

Nielsen, F., and Nielsen, F. (2016). Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, (pp. 195–211).

Nino Rondon, C., Castro Casadiego, S., Guevara, D., Castellano Carvajal, D., and Medina Delgado, B. (2022). An approach to edge detection in medical imaging through histogram analysis and morphological gradient. *Ingenieria y Competitividad*, *24*, 1–18.

Noorbehbahani, F., Salehi, F., and Zadeh, R. (2019). A systematic mapping study on gamification applied to e-marketing. *Journal of Research in Interactive Marketing*, *13*(3), 392–410.

Odhiambo, J., Weke, P., and Ngare, P. (2020). Modeling kenyan economic impact of corona virus in kenya using discrete-time markov chains. *Journal of Finance and Economics*, *8*, 80–85.

Okwuashi, O., and Ndehedehe, C. (2020). Integrating machine learning with markov chain and cellular automata models for modelling urban land use change. *Remote Sensing Applications Society and Environment*.

Olivier, L., Botha, S., and Craig, I. (2020). Optimized lockdown strategies for curbing the spread of covid-19: A south african case study. *IEEE Access*, *8*, 205755–205765.

Onder, I., and Berbekova, A. (2021). Web analytics: more than website performance evaluation? *International Journal of Tourism Cities*, *8*, 603–615.

Ortiz-Posadas, M. R. (2020). *Pattern Recognition Techniques Applied to Biomedical Problems*. Springer.

Pargent, F., Pfisterer, F., Thomas, J., and Bischl, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*, *37*(5), 2671–2692.

Patel, S., and Upadhyay, S. (2020). Euclidean distance based feature ranking and subset selection for bearing fault diagnosis. *Expert Systems with Applications*, *154*, 113400.

Pereira, R. C., Rodrigues, P., Figueiredo, M., and Henriques Abreu, P. (2023). *Automatic Delta-Adjustment Method Applied to Missing Not At Random Imputation*. Springer.

Perski, O., Blandford, A., West, R., and Michie, S. (2016). Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Translational Behavioral Medicine*, *7*(2), 254–267.

Picardi, D. (2022). Conditional probabilities and bayesian theorem in the study of animal disease introduction. (a didactic approach in environmental risk analysis). *MethodsX*, *9*, 101870.

Pirvu, C. M., and Anghel, A. (2019). Predicting next shopping stage using google analytics data for e-commerce applications. *arXiv*, *1*(1), 1–11.

Pitts, S., and Robinson, K. (2016). Dropping in and dropping out: experiences of sustaining and ceasing amateur participation in classical music. *British Journal of Music Education*, *33*, 1–20.

Porsche, L., Sucha, L., and Martinek, J. (2022). The potential of google analytics for tracking the reading behavior in web books. *Digital Library Perspectives*, *38*(1), 1–10.

Poulos, M., Korfiatis, N., and Papavlassopoulos, S. (2020). Assessing stationarity in web analytics: A study of bounce rates. *Expert Systems*, *37*(3), 12502.

Rahman, A., Dash, S., Luhach, A., Chilamkurti, N., Baek, S., and Nam, Y. (2019). A neuro-fuzzy approach for user behaviour classification and prediction. *Journal of Cloud Computing Advances Systems and Applications*, *8*, 1–15.

Raman, R., Gupta, N., and Jeppu, Y. (2023). Framework for formal verification of machine learning based complex system-of-systems. *INSIGHT*, *26*, 91–102.

Richiello, M., Mawdsely, G., and Gutman, L. (2021). Using the behaviour change wheel to identify barriers and enablers to the delivery of webchat counselling for young people. *Counselling and Psychotherapy Research*, *22*(1), 130–139.

Rita, P., Oliveira, T., and Farisa, A. (2019). The impact of e-service quality and customer satisfaction on customer behavior in online shopping. *Heliyon*, *5*(10), e02690.

Rodda, S., Hing, N., Hodgins, D., Cheetham, A., Dickins, M., and Lubman, D. (2018). Behaviour change strategies for problem gambling: an analysis of online posts. *International Gambling Studies*, (pp. 1–19).

Rojas, J., Riveros, A., Mejia, A., and Acosta-Prado, J. C. (2022). Positioning and web traffic of colombian banking establishments. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*, 1473–1492.

Rosqa, J. F., and Ati, M. (2022). Analysis of google analytics at elzatta's website in 2022. *e-Proceeding of Applied Science*, *6*(1), 2719–2726.

Santos, T., Silva, I., and Bessani, M. (2022). Evolving dynamic bayesian networks by an analytical threshold for dealing with data imputation in time series dataset. *Big Data Research*, *28*, 100316.

Santos, T., Teixeira, A., Terra, F., Moreira, L., and Toma, R. (2021). Detecting desertification in different years and rainfall regimes by 2d scatter plot. *Revista Ciencia Agronomica*, *52*.

Schmitt, L., Haupenthal, I., and Ahmed, F. (2021). Website design and trust elements: A/b testing on a start-up's website. *Enterprise Research Innovation*, *7*, 170–180.

Schule, M., Lang, H., Springer, M., Kemper, A., Neumann, T., and Gunnemann, S. (2021). In-database machine learning with SQL on GPUs. *Association for Computing Machinery*, (pp. 25–36).

Selvanathan, S. P., Cuce, E., and Sudhakar, K. (2021). A perspective of covid 19 impact on global economy, energy and environment. *International Journal of Sustainable Engineering*, *14*, 1–16.

Semeradova, T., and Weinlich, P. (2020). *Using Google Analytics to examine the website traffic*. Cham: Springer International Publishing.

Shi, Y. (2022). *Advances in Big Data Analytics: Theory, Algorithms and Practices*. Singapore: Springer.

Shijitha, R., Karthigaikumar, P., and Paul, A. (2022). Data warehouse design for big data in academia. *Computers, Materials and Continua*, *71*, 979–992.

Shrifan, N., Akbar, M. F., and Mat Isa, N. A. (2021). An adaptive outlier removal aided k-means clustering algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*.

Sinaga, K. P., and Yang, M.-S. (2020). Unsupervised k-means clustering algorithm. *IEEE Access*, *8*, 80716–80727.

Singh, V., Pencina, M., Einstein, A., Liang, J., Berman, D., and Slomka, P. (2021). Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific Reports*, *11*.

Soltaninejad, M., Zhang, L., Lambrou, T., Yang, G., Allinson, N., and Ye, X. (2018). Mri brain tumor segmentation and patient survival prediction using random forests and fully convolutional networks. *Lecture Notes in Computer Science*, *10670*.

Sood, S. (2022). *Leveraging web analytics for optimizing digital marketing strategies*. Taylor and Francis Group.

Spedicato, G. A., Kang, T. S., Bhargav, S., and Spedicato, M. G. A. (2015). Package 'markovchain'.

Stacke, K., Eilertsen, G., Unger, J., and Lundstrom, C. (2020). Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, *PP*, 1–1.

Stelian, N., and Stoicu-Tivadar, L. (2020). Evaluating interactivity in vr healthcare applications using analytics. *Studies in health technology and informatics*, *272*, 225–228.

Sullivan, J. H., Warkentin, M., and Wallace, L. (2021). So many ways for assessing outliers: What really works and does it matter? *Journal of Business Research*, *132*, 530–543.

Takai, K., and Hayashi, K. (2023). Model selection with missing data embedded in missing-at-random data. *Stats*, *6*, 495–505.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. (2020). Measuring robustness to natural distribution shifts in image classification. *ArXiv*, (pp. 1–69).

Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical Sonography*, *6*, 35–39.

Thirumalai, C. S., Manickam, V., and Balaji, R. (2017). Data analysis using box and whisker plot for lung cancer. *2017 Innovations in Power and Advanced Computing Technologies*, (pp. 1–6).

Thiruvengadam, G., Marappa, L., and Ramanujam, R. (2021). A study of factors affecting the length of hospital stay of covid-19 patients by cox-proportional hazard model in a south indian tertiary care hospital. *Journal of Primary Care and Community Health*, *12*, 215013272110002.

Thushara, Y., and Vamanan, R. (2016). A study of web mining application on e-commerce using google analytics tool. *International Journal of Computer Applications*, *149*, 975–8887.

Tien, J. M. (2017). Internet of things, real-time decision making, and artificial intelligence. *Annals of Data Science*, *4*(2), 149–178.

Tinguely, R., Montes, K., Rea, C., Sweeney, R., and Granetz, R. (2019). An application of survival analysis to disruption prediction via random forests. *Plasma Physics and Controlled Fusion*, *61*(9), 095009.

Trieu, V.-H., Burton-Jones, A., Green, P., and Cockcroft, S. (2022). Applying and extending the theory of effective use in a business intelligence context. *MIS Quarterly*, *46*, 645–678.

Tunio, M. N., Shaikh, E., Katper, N., and Brahmi, M. (2023). Nascent entrepreneurs and challenges in the digital market in developing countries. *International Journal of Public Sector Performance Management*, *12*, 140–153.

Uba, G., Zandam, N. D., Mansur, A., and Shukor, M. Y. (2021). Outlier and normality testing of the residuals for the morgan-mercer-flodin (MMF) model used for modelling the total number of covid-19 cases for Brazil. *Bioremediation Science and Technology Research*, *9*(1), 13–19.

Uli, P. D. C. F., and Laksmidewi, D. (2023). Web design and web brand image of online travel agent: Its effect on purchase intention mediated by trust. *International Journal of Science and Society*, *5*(1), 399–411.

Urban, C., and Mine, A. (2021). A review of formal methods applied to machine learning. *arXiv*.

Uzun Ozsahin, D., Mustapha, M., Mubarak, A., Ameen, Z., and Uzun, B. (2022). Impact of outliers and dimensionality reduction on the performance of predictive models for medical disease diagnosis. *2022 International Conference on Artificial Intelligence in Everything*, (pp. 79–86).

Van Gassen, S., Gaudilliere, B., Angst, M. S., Saeys, Y., and Aghaeepour, N. (2020). Cytonorm: a normalization algorithm for cytometry data. *Cytometry Part A*, *97*(3), 268–278.

Vazifehdan, M., Moattar, M., and Jalali, M. (2019). A hybrid bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction. *Journal of King Saud University - Computer and Information Sciences*, *31*.

Vecchione, A., Brown, D., Allen, E., and Baschnagel, A. (2016). Tracking user behavior with google analytics events on an academic library web site. *Journal of Web Librarianship*, *10*, 1–14.

Vermeer, S., and Trilling, D. (2020). Toward a better understanding of news user journeys: A markov chain approach. *Journalism Studies*, *21*, 879–894.

Vomlel, J., Kratochvil, V., and Kratochvil, F. (2023). Structural learning of mixed noisy-or bayesian networks. *International Journal of Approximate Reasoning*, *161*, 108990.

Waggoner, P., Kennedy, R., and Clifford, S. (2019). Detecting fraud in online surveys by tracing, scoring, and visualizing ip addresses. *Journal of Open Source Software*, *4*, 1285.

Walsh, D., Hall, M., Clough, P., and Foster, J. (2020). Characterising online museum users: A study of the national museums liverpool museum website. *International Journal on Digital Libraries*, *21*.

Wang, L., and Law, R. (2019). Relationship between hotels' website quality and consumers' booking intentions with internet experience as moderator. *Journal of China Tourism Research*, *16*, 1–21.

Wang, S., and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *The International Biometric Society*, *64*(2), 440–448.

Wang, X., Li, Y., Cai, Z., and Liu, H. (2021). Beauty matters: reducing bounce rate by aesthetics of experience product portal page. *Industrial Management and Data Systems*, *121*(8), 1848–1870.

Wang, Y., Cai, J., Zhang, D., Chen, X., and Wang, Y. (2022). Nonlinear correction for fringe projection profilometry with shifted-phase histogram equalization. *IEEE Transactions on Instrumentation and Measurement*, *71*, 5005509.

Weber, J. (2015). *Practical Google Analytics and Google Tag Manager for Developers*. Apress.

Wilinski, A. (2019). Time series modelling and forecasting based on a markov chain with changing transition matrices. *Expert Systems with Applications*, *133*.

Winklbauer, S., and Horner, R. (2022). Austria: Data protection authority sees use of google analytics as unlawful transfer of data. *Computer Law Review International*, *23*(1), 30–32.

Wongvibulsin, S., Wu, K., and Zeger, S. (2019). Clinical risk prediction with random forests for survival, longitudinal, and multivariate (rf-slam) data analysis. *BMC Medical Research Methodology*, *20*, 1.

Wu, H., Cao, Y., Wei, H., and Tian, Z. (2021). Face recognition based on haar like and euclidean distance. *Journal of Physics: Conference Series*, *1813*, 012036.

Xiong, Y., Liu, X., Lan, L.-C., You, Y., and Hsieh, C.-J. (2020). How much progress have we made in neural network training? a new evaluation protocol for benchmarking optimizers. *ArXiv*, (pp. 1–16).

Xu, J., Zhang, Y., and Miao, D. (2019). Three-way confusion matrix for classification: A measure driven view. *Information Sciences*, *507*.

Xuhua, H., Ocloo, E., Akaba, S., and Worwui-Brown, D. (2019). Effects of business to business e-commerce adoption on competitive advantage of small and medium-sized manufacturing enterprises. *Economics and Sociology*, *12*.

Yao, Y. (2023). Bayesian network model structure based on binary evolutionary algorithm. *PeerJ Computer Science*, *9*, e1466.

Ye, C., Wang, H., Lu, W., and Li, J. (2019). Effective bayesian-network-based missing value imputation enhanced by crowdsourcing. *Knowledge-Based Systems*, *190*.

Zhou, Y., Shi, J., Stein, R., Liu, X., Baldassano, R., Forrest, C., Chen, Y., and Huang, J. (2023). Missing data matter: an empirical evaluation of the impacts of missing ehr data in comparative effectiveness research. *Journal of the American Medical Informatics Association*, *30*.