Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/12S106

A tiered approach to Markov models when future events are not independent of the past: an application in web analytics

Judah Soobramoney[*], *Retius Chifurira, Knowledge Chinhamu and Temesgen Zewotir*

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, South Africa

Received: 27 Jul. 2023, Revised: 15 Oct. 2023, Accepted: 15 Oct. 2023
Published online: 1 Dec. 2023

Abstract: The internet has become a digital marketplace that offers goods and services globally. Thereby, compelling enterprises to optimize their websites and online strategies. This paper sought to employ Markov chain models to predict the most likely next webpage viewed. A website comprises several pages (such as "Home", "About-Us") and visitors would transition from one page to another by clicking on respective links. The study was conducted on the website of a South African engineering and engineering training company "TEKmation". The transition probabilities therefore represent the likelihood of moving to a certain webpage, given that the visitor is on a specific webpage all within the studied website. However, a key Markov chain assumption is that the "next state" is solely dependent on the "present state" and independent of "previous states" ("memoryless"). However, according to chi-squared tests on the observed data, the "future" state has shown dependence on "previous" states. And this was due to a visitor being less likely to re-visit a page again relative to the likelihood of visiting an unseen page within the visit. The aim of this study was to explore a tiered approached to the Markov models to minimize the impact of the "memoryless" assumption. The study further split each visit into a tier one portion (which represented the first two viewed pages of the visit) and a tier two portion (which represented the third or more pages viewed). The tiered approach (accuracy = 62%) fitted the data a lot better than the standard Markov model (accuracy = 53%). It was also observed that the tiered model had on average more accurately predicted the drop-off events (the movement from the "current state" to exiting the website). Thereby, in conclusion, the tiered Markov models proved to reduce the "memoryless" assumption on the studied data.

Keywords: chi-squared test, Markov chains, transition probabilities, web analytics

1 Introduction

The world wide web has become a competitive global marketplace. In developed countries, online shopping is a norm whilst in developing countries, online shopping is becoming increasingly more popular [1,2]. Consumers have access to goods and services that are available globally [3,4]. The wellbeing of Corporates are thus materially influenced by the corporate's website. A website that is attractive, convincing, and user-friendly is more likely to onboard customers and ultimately maintain or increase revenue [5]. The data scientist community thereby, seek to optimize web design and user-experience (UX) to ultimately maximize profits. Through the use of web tracking tools, such as "GoogleAnalytics", data scientists have access to visitor information such as the device used (mobile/desktop/tablet), operating system, browser used (Chrome, Firefox, and others), the brand of the device, the geo-location of the device, the number of times the visitor has entered the website previously, the duration on the website, the pages viewed and much more [6].
This paper sought to predict the next webpage that a visitor would most likely view whilst on a visit. By construct, a website is made up of webpages where a visitor entering the website would navigate across several webpages by clicking on links. For example, a person would first enter onto the "Home" page and thereafter click and enter the "Contact-us" webpage and thereafter exit the website. Whilst there are several machine learning models that could be employed, the authors have investigated the use of Markov chains. 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 webpage, and push pop-ups to guide the page path to the desired webpages. Although more

* Corresponding author e-mail: judahsoobramoney@gmail.com

sophisticated machine learning algorithms are noted to yield high prediction accuracies (such as Random Forests) [7], the implementation of these models were a challenge on the studied website and would potentially hinder browsing speed. However, a key assumption of Markov chains is that the future state is only dependent on the present state and not previous states ("memorylessness") 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 seen. Therefore, to improve the accuracy of the Markov transition probabilities in the context of the studied website, a tiered Markov chain model was proposed. 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.

1.1 Related work

The paragraphs below discusses recent applications of web behaviour prediction and recent applications of 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 [8]. 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 [9]. Rahman et al. (2019) employed a neuro-fuzzy approach to predict online behaviour using people's browsing interests and observing suspicious activities (such as security and privacy) derived from their internet trail. The proposed model was found to be promising in terms of the classification and prediction accuracy [10]. 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 [11]. Vermeer and Thrilling (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 [12]. 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 level of accuracy and proved to be a robust method for modelling urban change [13]. Given the recent literature reviewed herein, none of the applications attempt to address cases where the memoryless property of the Markov model may not realistically apply.

1.2 Research highlights

On the observed website, a predictive model was required to determine the most likely next webpage a visitor would select. The predictive model was required to be simplistic to allow easy implementation within the underlying code behind the website to ultimately lure visitors onto the more 'important' webpages. Whilst Markov models were considered to be simple enough (due to the transition probability matrix), the 'memoryless' assumption of Markov models were not ideal. The purpose of the study aimed to evaluate a tiered approach to employing Markov models to address the 'memoryless' assumption.

Section 2 of this paper discusses the theoretical framework of Markov chains, section 3 discusses the data source and data processing, section 4 explains the derivation of the tiers and presents the tiered Markov models. And section 5 discusses the results of the models, the limitations, future work and describes the use case of the proposed models.

2 Markov methodology

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

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:

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

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 [14, 15].

2.2 Transition matrix

Assume a state space ($T = 1, \ldots, d$) 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 [16]:

$$\begin{pmatrix} T^{00} & \ldots & T^{0d} \\ \ldots & \ldots & \ldots \\ T^{d0} & \ldots & T^{dd} \end{pmatrix}. \tag{2}$$

2.3 Absorbing state

An absorbing state refers to a state that can be reached from any other state. But once in an absorbing state, the random variable cannot leave that state [17].

In the context of this study, where a state refers to the "webpage" being viewed, at any given state a visitor could chose to subsequently leave the website (drop-off). However, once a person has dropped-off , the visitor could not transition to any other states afterwards. Should the "dropped-off" visitor re-enter the website, it would be treated as another visit.

2.4 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 [18]:

$$P(X_{t+1} = a \mid X_t = b) = P(X_1 = a \mid X_0 = b). \tag{3}$$

3 Markov data

The data sourced for this analysis was derived from the Google Analytics web tracking tool. The investigation was conducted for a South African corporate in the engineering and engineering training sector (TEKmation) and thus the behaviour of online users and corresponding data would be specific to such industry. The tracking tool supplies 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 1 below depicts a typical visit journey.

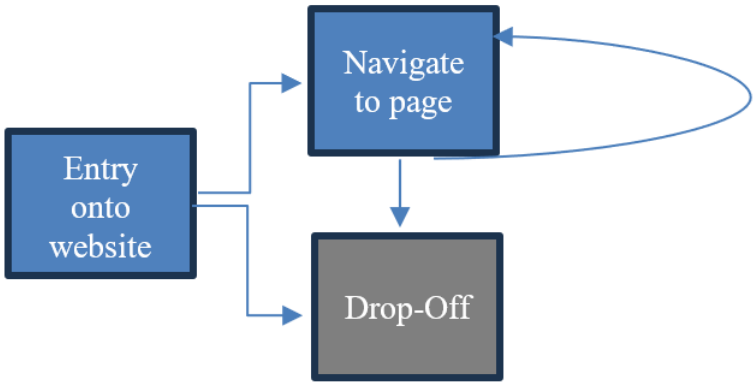

Fig. 1: Online visit page path,

The "Entry onto website" block of Figure 1, represents the phase a person would enter the website on a given webpage. It is possible for a person to enter the website on any given webpage (for example, some may enter and land on the "Home" page first, others may land on the "Contact-us" page first). Therefore, the order of pages viewed would vary across the visitors. Whilst most links would route to the "Home" page first, web browsers (such as "Google") typically yield a few options to the person prior to entering the website allowing visitors to land on several pages of the website (such as 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 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 page. For example, all of the detailed "Courses" pages that elaborated on educational courses offered by the corporate have been rolled up to the state: "Courses". The root pages can be viewed in Figure 2 for context.

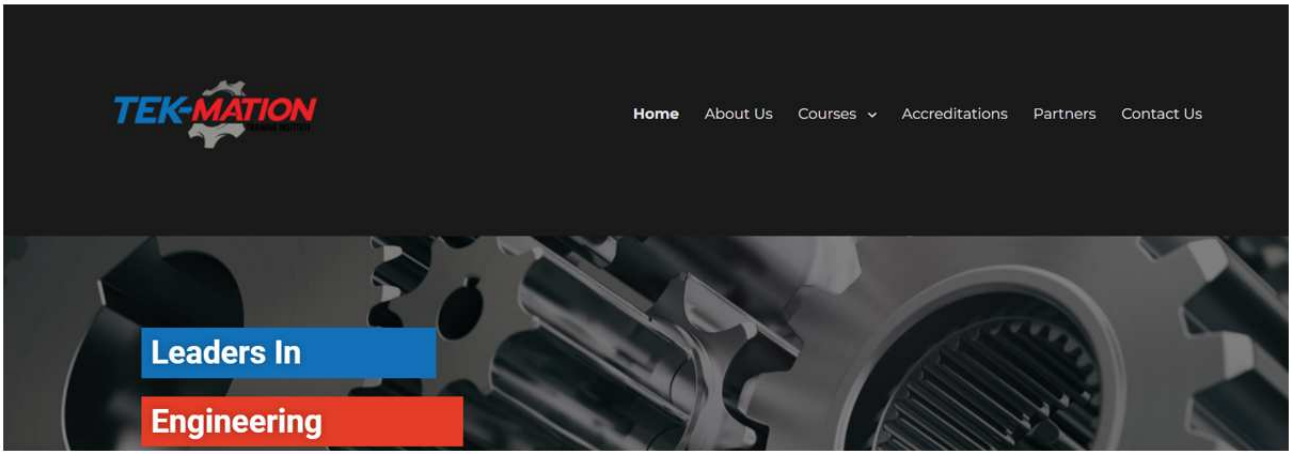

Fig. 2: Studied website extract,

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 “DropOff”.

4 Markov results

This section, firstly, tests the assumption of “memoryless” webpage transitioning through chi-squared tests. 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 webpage 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” [19]).

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 would less likely 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 below (Figure 3) indicates that on studied website, on a given webpage (w_0) the transition to the next webpage (w_1) is dependent on the previous webpage (w_{-1}).

```
>
> #------------------- Chi-Squared Test -----------------
>
> chisq.test(d$PrevState, d$NextState, correct=FALSE)

        Pearson's Chi-squared test

data:  d$PrevState and d$NextState
X-squared = 1856.1, df = 36, p-value < 2.2e-16
```

Fig. 3: Previous state and next state Chi-squared test for dependence,

Since p-value 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.

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 4 depicts the distribution the total count of webpages viewed on the studied website.

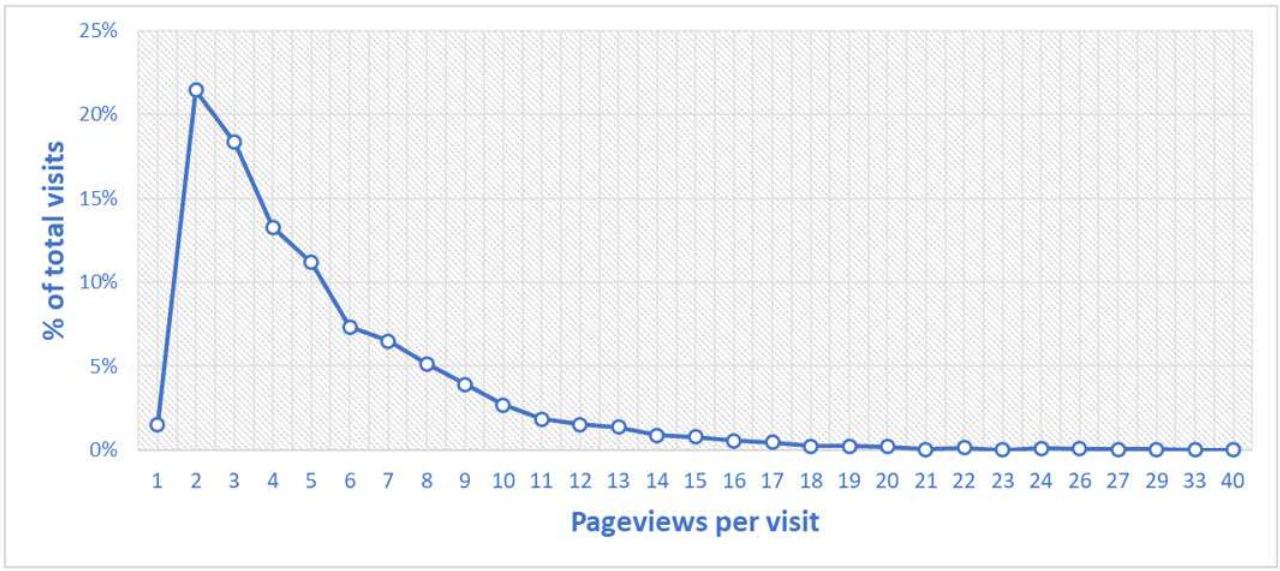

Fig. 4: Distribution of the number of pages viewed per visit,

Using the distribution of page-views, 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 page. And the second tier would hold the transition probabilities of the people who were on their third or more webpage. Whilst adding more tiers may more accurately enhance the prediction of the next state, the authors strived to keep the process as simple as possible to allow for practical application.

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 memory-less 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 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 1 below quantifies the transitional probabilities of the tier 1 Markov model.

Table 1: Tier 1 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 1 represent the current state (current webpage viewed) given that the current visit has just started (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 1 transition probabilities, a visitor has a 20% probability of

navigating to the "Contact" page from the "About" page after the first 2 pages are viewed. Figure 5 depicts the Tier one Markov chains as represented by the transition probabilities.

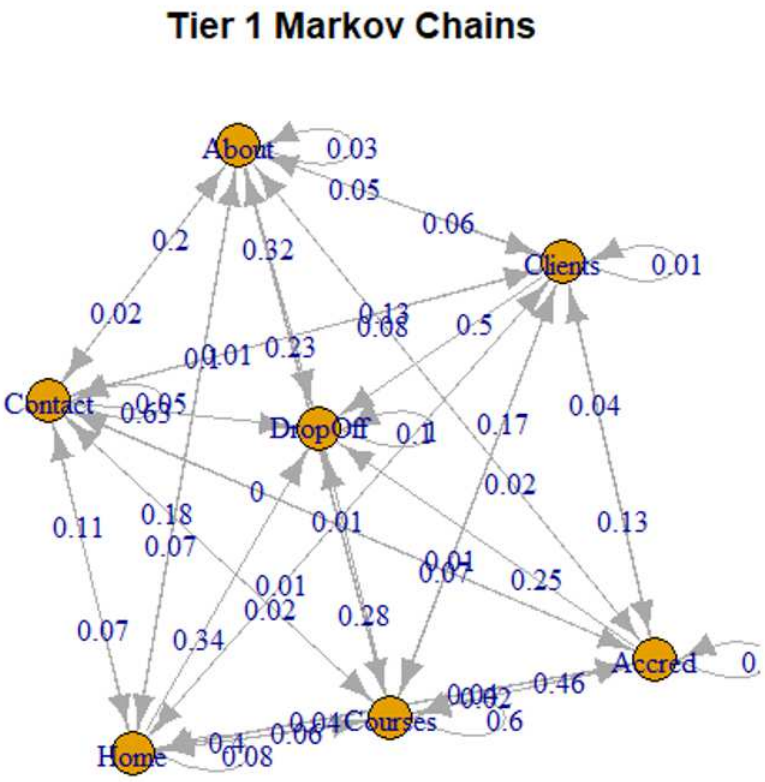

Fig. 5: Tier 1 Markov chains,

The Tier 1 Markov chains presented in Figure 5, depicts the seven current states and quantify the transition probability of moving to the next state. It can also be seen that the state "DropOff" is an absorbing state meaning that once the current state is at "DropOff" there will be no subsequent state. The state "DropOff" 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 would be treated as a new visit.

Table 2 below presents the Tier 2 transition probabilities on the studied website. These probabilities represent the likelihood that a visitor would transition to another webpage given that the current page is the third or more webpage being viewed by the visitor.

Table 2: Tier 2 transition probabilities,

Tier 2	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 2 represent the current state, and the columns represent the next state. According to the Tier 2 transition probabilities, as per Table 2, there is a 13% chance of transitioning to the "Contact" webpage from the "About" webpage. Figure 6 depicts the Tier 2 Markov chains.

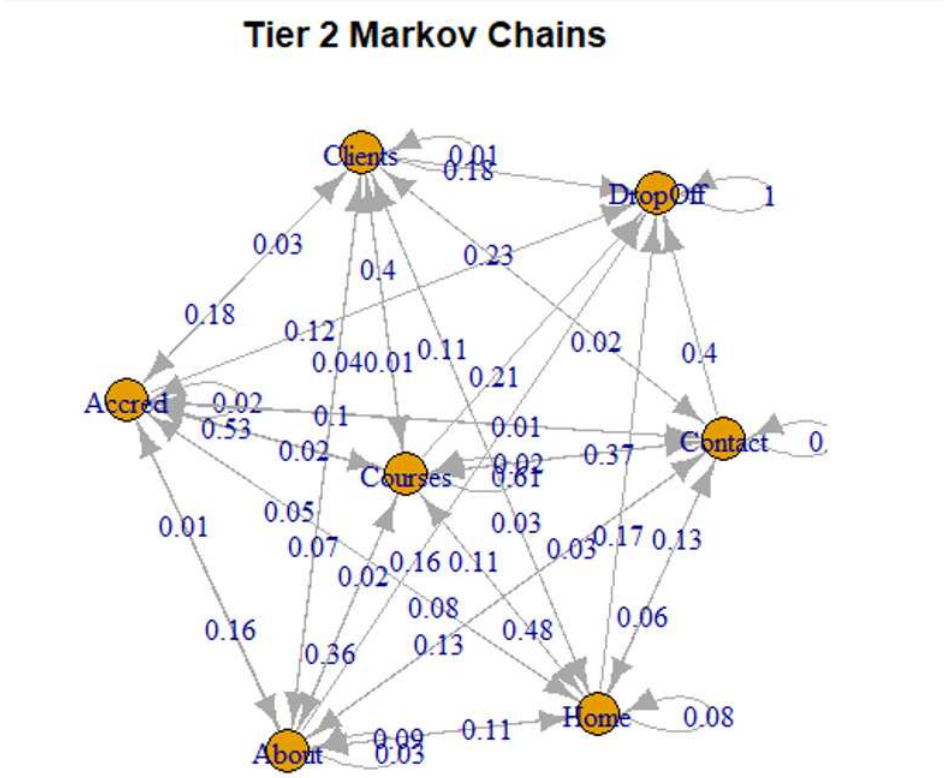

Fig. 6: Tier 2 Markov chains,

Similar to the Tier 1 Markov chain model, Figure 6 depicts the seven current states and quantify the transition probability of moving to the next state.

5 Concluding remarks

5.1 Discussion

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 minimize 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.

Upon studying the Tier 1 and Tier 2 transition probabilities, it was evident that the probability distributions were more discriminant. Tier 1 highlighted that when visitors were on the "Clients" and "Contact-Us" webpages early on in their journey, they held a high likelihood of dropping off. Table 3 quantifies the average probability of moving to the respective states.

Table 3: Average transition probability of the 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 3, that the tiered models have resulted in greater differentiation than 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 3, the Tier 1 model would more accurately predict drop-offs. Whilst the Tier 2 model would more accurately predict movement to other pages relative to the Tier one model.
Whilst 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.

5.2 Limitations and future work

A key limitation of the study was the 'memoryless' assumption of Markov models. Therefore, whilst other Bayesian models such as Bayesian network models would yield higher prediction accuracies, the model needed to be as simple as possible to implement. Future work could explore further splitting the Tier 2 model into finer Markov models and explore methods to determine the optimal number of tiers. Whilst there may be various techniques to determine the optimal number of tiers, the elementary technique of 'test-and-learn' could be initially tested where the prediction accuracies of a one-tier, two-tier, five-tier, n^{th} tier be constructed to identify the most accurate tier structure that best predicts a visitor's next action. However, it would be important to note that the more tiers employed, the higher the complexity and the lower the sample size may be. Furthermore, the optimal number of clusters will be specific to a particular website. However, within the application of this study, the solution had to be as light-weight as possible, thereby a two-tier solution was employed.
On the studied website, the Markov transition probabilities would be coded into the website infrastructure to firstly, predict visitor drop-offs and push prompts to prolong the visit on the website. Secondly, the board of directors claim that certain webpages are more important than others and would therefore attempt to redesign 'less-important' webpages to lure visitors onto the more important webpages. These changes would first be tested through an AB test and thereafter rolled out.

Declarations

–Acknowledgments: We thank the Editor-in-Chief and the reviewers for the valuable feedback to improve the quality and impact of this paper.
–Funding: No funding was received.
–Conflict of interest/Competing interests: On behalf of all authors, the corresponding author states that there is no conflict of interest.

–Ethics approval: Not applicable.
–Availability of data and materials: None. Unfortunately, the data belongs to the owner of the website.
–Code availability: R code available upon request.
–Authors' contributions: JS: (50%) methodology, writing and preparation. RC: (20%), conceptualization, investigation, supervision and editing. KC: (20%), conceptualization, investigation, supervision and editing. TZ: (10%) high-level supervision.

References

[1] M. Khan, S.S. Zubair and M. Malik, An assessment of e-service quality, e-satisfaction and e-loyalty: Case of online shopping in Pakistan, *South Asian Journal of Business Studies*, **8**, 1 (2019).
[2] M.N. Tunio, E. Shaikh, N. Katper and M. Brahmi, Nascent entrepreneurs and challenges in the digital market in developing countries, *International Journal of Public Sector Performance Management*, **12**, 140-153 (2023).
[3] J. Hu and A. Haddud, Exploring the Impact of Globalization and Technology on Supply Chain Management: A Case of International E-Commerce Business, *ISBN = 9781799809463*, 1353-1376 (2020).
[4] Y. Luo, New OLI advantages in digital globalization, *International Business Review*, **30**, 101797 (2021).
[5] L. Wang and R. Law, Relationship between Hotels' Website Quality and Consumers' Booking Intentions with Internet Experience as Moderator, *Journal of China Tourism Research*, **16**, 1-21 (2019).
[6] V. Kumar and G. Ogunmola, Web Analytics for Knowledge Creation: A Systematic Review of Tools, Techniques, and Practices, *International Journal of Cyber Behavior, Psychology and Learning*, **10**, 1-14 (2019).
[7] M.R. Dileep, A.V. Navaneeth and M. Abhishek, A Novel Approach for Credit Card Fraud Detection using Decision Tree and Random Forest Algorithms, *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, 1025-1028 (2021).
[8] D. Koehn, S. Lessmann and M. Schaal, Predicting Online Shopping Behaviour from Clickstream Data using Deep Learning, *Expert Systems with Applications*, **150**, 113342 (2020).
[9] P. Nagaraj, V. Muneeswaran, A. Dharanidharan, M. Aakash, K. Balananthanan, and C. Rajkumar, E-Commerce Customer Churn Prediction Scheme Based on Customer Behaviour Using Machine Learning, *International Conference on Computer Communication and Informatics (ICCCI)*, 1-6 (2023).
[10] A. Rahman, S, Dash, A. Luhach, N. Chilamkurti, S. Baek and Y. Nam, A Neuro-fuzzy approach for user behaviour classification and prediction, *Journal of Cloud Computing Advances Systems and Applications*, **8**, 1-15 (2019).
[11] Z. Jia, Z. Zhou, H. Zhang, B. Li and Y. Zhang, Forecast of coal consumption in Gansu Province based on Grey-Markov chain model, *Energy*, **199**, 117444 (2020).
[12] S. Vermeer, and D. Thrilling, Toward a Better Understanding of News User Journeys: A Markov Chain Approach, *Journalism Studies*, **21**, 879-894 (2020).
[13] O. Okwuashi, and C. Ndehedehe, Integrating machine learning with Markov chain and cellular automata models for modelling urban land use change, *Remote Sensing Applications Society and Environment*, **1**, 1 (2020).
[14] J. Odhiambo, P. Weke and P. Ngare, Modeling Kenyan Economic Impact of Corona Virus in Kenya Using Discrete-Time Markov Chains, *10.12691/jfe-8-2-5*, **8**, 80-85 (2020).
[15] J. Cortes, S.K. El-Labany, A. Navarro-Quiles, M. Selim, and H. Slama, 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**, 1 (2020).
[16] A. Wilinski, Time Series Modelling and Forecasting Based on a Markov Chain with Changing Transition Matrices, *Expert Systems with Applications*, **133**, 1 (2019).
[17] Y. Li and W. Ji, Understanding The Dynamics Of Information Flow During Disaster Response Using Absorbing Markov Chains, *2020 Winter Simulation Conference (WSC)*, 2526-2535 (2020).
[18] P. Mahmoudi and A. Rigi, Probabilistic Prediction of Drought in Iran Using Homogenous and Nonhomogenous Markov Chains, *Journal of Hydrologic Engineering*, **28**, 1 (2023).
[19] , G.A. Spedicato, T.S. Kang, S. Bhargav, M.G.A. Spedicato, Package 'markovchain', *CRAN*, (2015).