Received: 30 December 2019 | Revised: 11 June 2020 | Accepted: 12 June 2020

DOI: 10.1002/asmb.2563

SPECIAL ISSUE PAPER

WILEY

Topological data analysis in digital marketing

Choudur Lakshminarayan[1,2] | **Mingzhang Yin**[1]

[1]Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, Texas, USA

[2]Teradata Labs, Austin, Texas, USA

Correspondence
Choudur Lakshminarayan, Teradata Labs, Austin, TX 78759.
Email: choudur.lakshminarayan@teradata.com

Abstract

The ubiquitous internet is a multipurpose platform for finding information, an avenue for social interaction, and a primary customer touch-point as a marketplace to conduct e-commerce. The digital footprints of browsers are a rich source of data to drive sales. We use clickstreams (clicks) to track the evolution of session-level customer browsing for modeling. We apply Markov chains (MC) to calculate probabilities of page-level transitions from which relevant topological features (*persistence diagrams*) are extracted to determine optimal points (URL pages) for marketing intervention. We use topological summaries (*silhouettes, landscapes*) to distinguish the *buyers* and *nonbuyers* to determine the likelihood of conversion of active user sessions. Separately, we model browsing patterns via Markov chain theory to predict users' propensity to buy within a session. Extensive analysis of data applied to a large commercial website demonstrates that the proposed approaches are useful predictors of user behavior and intent. Utilizing computational topology in digital marketing holds tremendous promise. We demonstrate the utility of topological data analysis combined with MC and present its merits and disadvantages.

KEYWORDS

clickstreams, computational topology, landscapes, Markov chains, persistence diagrams, silhouettes

1 | MOTIVATION AND RELATED LITERATURE

The holy grail of digital marketing is reaching the right customer in real time with the right offer based on browsing history. It has been reported that for every one dollar spent, 92 cents are spent in acquiring a customer. Needless to say, converting leads into buyers is a priority. Sadly, the currently available mother lode of online customer information has not helped to move conversion beyond the 2% to 3% range. In dealing with web browsing behaviors, we want to develop quantitative methods to distinguish between two types of users: *buyers* and *nonbuyers*. The mainstay of consumer analytics is confined to counting discrete web events, such as number of visitors, referrals, abandonment rate, exit pages, pages viewed, and average page duration, to correlate to an outcome (*conversion*).[1-3] Clearly, modeling *conversion* as a function of aggregated web events is fraught with ambiguities due to smoothing over customer interactions with the web. Parametric models such as hierarchical Bayes models, multinomial probit models, hidden Markov chain models, and vector autoregressive models have been applied.

Alternately, when viewing an enterprise website as a coherent network of related pages connected by hyperlinks, a visitor's session-level sequence of clicks is a navigation through a set of pages resulting in either purchase or abandonment. The presumed interrelationships between pages in visitor sessions readily lend themselves to modeling by a class of stochastic processes known as Markov chains. Markov chains in web usage mining have been studied extensively.[4-8] This has included taking a Markov chain approach to predict conversion based on page transitions from

session navigations, thereby establishing a closer link to *conversion*. However, in these approaches, the intrasession patterns are neglected. Active web sessions demonstrate unique patterns (shapes) that can be exploited to distinguish sequences leading to conversion or abandonment. Topological data analysis (TDA) posits that data have shape and shapes help distinguish users' browsing patterns. While the application of machine learning in clickstream analytics seems straightforward, idiosyncratic customer browsing, arbitrary session definitions, and page instrumentation make modeling difficult. This is further exacerbated by low conversion rates for certain consumer electronic products. In the ensuing sections, we model clickstreams by traditional Markov chains, and by computational topology. The page transition matrices (TPM) used in Markov modeling are leveraged to derive topological features to tease out intrasession navigational patterns.

Early work on clickstreams has relied on web usage data collected from media agencies that recruit and track a representative sample of volunteers over extended periods, confined to commodity products such as books. Clickstreams have been analyzed using association rules to find sets of pages within sessions that exceed a minimum *support* threshold. The method of sequential patterns is used to find intersession patterns such that marketers can predict future user patterns. Furthermore, user profiles constructed from clickstreams are used to classify users into different classes (*buyer*, *nonbuyer*) using classifiers such as decision trees, naive Bayes, support vector machines, and k-nearest neighbors. Dynamic multinomial probit models to introduce memory were proposed by Montgomery et al.[1] In an early paper on web usage mining, Srivastava et al[2] outline quantitative approaches to clickstreams. Moe and Fader[3] study the relationship between visiting frequency (interarrival time) and propensity to buy using exponential-gamma distributions to account for variable arrival rates among individuals in the panel. Conversion is a primary key performance index (KPI) in digital marketing. Eirinaki et al[7] have applied Markov chain approaches for *page ranking*. The benchmark conversion prediction rate for commodity products is reported to be 7% without clickstream path information, while it is 40% within six page clicks. We will see that our models applied to consumer electronic products deliver *recall* in excess of 80% and a *false positive rate* of less than 10% within five clicks.

2 | DATA SETS, DESCRIPTION, AND DATA PREPARATION

We tested our proposed methods using consumer electronics products (desktops, notebooks, printers, and supplies). We present results only from models applied to electronic peripherals known as *supplies* (results from other products will be made available as needed). The data set for a given product is a table (T), as seen in Table 1, where each row consists of a session ID, page name (URL), and class label ($Buy = 1$, $No\ Buy = 0$). The data set consists of over two million unique sessions, a total of 5722 unique URL pages with an overall conversion rate of ∼1.5%. The average length of sessions that resulted in a purchase is ∼36 clicks with a standard deviation of 21 clicks (pages). The median session length is ∼25 clicks. For the nonbuyers, the average, standard deviation, and median (approximately) are 20 clicks, 16 clicks, and 11 clicks, respectively. The objective is to identify prospects who will convert based on their session-level page navigation. To do this, we divide the data set into two classes ($C_i, i = 1, 2$): buyers and nonbuyers. The *training* data is composed by obtaining 70% from ($C_i, i = 1, 2$) and the *testing* set is obtained by combining the remaining 30%. Class imbalance is a vexing problem in the real world, as evidenced by low conversion rates across all product lines. The nonbuyers tend to make up 98% to 99% because the population may include online visitors who only search, download drivers, or retrieve PDF documents. To reduce the presence of such sessions that have no intention of buying, we undersample the nonbuyers such that a ratio of 10 to 20 is maintained for classification purposes. Upon sampling, we collect all URL sequences from the *training* set to create transition probability matrices (TPM) for the two classes C_1 and C_2. Each session-level URL sequence in C_1 is decomposed into pairwise URLs $i \rightarrow j$, where the transitions $i \rightarrow j$ represent *From* and *To* pages occurring in sessions

TABLE 1 A sample user session

Session ID	Page	Class
3206114067796127468696417 50-4	us:welcome-home	0
3206114067796127468696417 50-4	us:sale:static:springsale	0
3206114067796127468696417 50-4	us:en-us:laptops notebook pc	0
3206114067796127468696417 50-4	us:laptops pavilion 15t-n200 notebook pc with windows 7	1

contiguously. For example, if a session consisted of page transitions $U_1 \to U_2 \to U_3$, the pairwise transitions would be $U_1 \to U_2$ and $U_2 \to U_3$. From the set of all $(i,j) \in C_1$, we compute transition probabilities p_{ij}. The pairwise probabilities p_{ij} form the TPM ($\mathbb{P}_1$) . Similarly, we generate the matrix $\mathbb{P}_2$ for class C_2. The two-class transition probability matrices (TPM) are the building blocks for constructing topological features to use in TDA and likelihood ratios in Markov models.

2.1 | Topology and data analysis

In science and engineering, certain phenomena lend themselves readily to embeddings in vector spaces with well-defined coordinates and characterizations using metrics such as *distances, norms, and angles*. In the case of web data, where the canonical unit is a URL page link, the definition of suitable metrics is not very clear. As we can only understand user behaviors only in a very coarse way due to idiosyncratic browsing, there is a need for alternative approaches that are insensitive to choice of the metrics and are coordinate-free. These issues inspired us to explore topology-based methods as an alternative. Unlike geometric methods, topology captures geometric properties of data in a way that is less sensitive to the actual choice of metrics. Topological data analysis replaces distance measures with the concept of infinite nearness of a point to a subset in the embedding space. This insensitivity to the metric is useful in studying clickstream data where the notion of a metric is arbitrary. Geometries of objects in topology are explained via summaries such as Čech complexes, Rips complexes, and Betti numbers. These summaries are related to *connectedness* in point-set clouds in topological spaces. Persistent homology[9] provides efficient algorithms to quantify the evolution of the topology of a family of nested topological spaces. To elaborate, given a real-valued function (f), persistent homology describes changes in the topology of lower level subsets $\{x : f(x) \leq a\}$ as a increases from $-\infty$ to $+\infty$. The persistence diagram encodes the changes via a multiset of points in a two-dimensional plane each corresponding to a *birth (b)* and *death (d)* of a homological feature for some interval $(-\infty, a]$. Features come on all scale levels and can be nested in more complicated relationships. One of the key challenges in persistent homology is to find a way to isolate the points of the persistence diagram representing the topological noise. Statistical methods enable estimation of persistence diagrams which approximate the true persistence diagram with high confidence. The sample-estimated persistence diagram is sufficient for practical purposes. Fasy et al[10] and Chazal et al[11] propose several statistical methods to construct confidence sets for persistence diagrams and other summary functions that allow us to separate topological signal from topological noise.

2.2 | Computational topology

Persistence homology postulates that patterns within a dynamic sequence can be decomposed topologically into transient and persistent segments. The *transients* correspond to noise, while *persistents* correspond to signal. In the following, we briefly discuss salient ideas from the field of persistent homology.

Given a real-valued function f, persistent homology[9] summarizes the topology of the lower level subsets $f^{-1}(-\infty, a_i]$ for increasing values of the parameter ($a_0 \leq a_2 \leq \ldots \leq a_n$). In other words, it generates a *filtration* $\mathcal{F}$ which is an indexed set of nested topological spaces, $\emptyset = \mathbb{S}_0 \subseteq \mathbb{S}_1 \subseteq \ldots \mathbb{S}_n$ with changing topological features. The multiset of points on a plane for each a_i generates connected components indexed by births (beginning) and deaths (end) denoted by b and d, respectively. The persistence diagram is a set of points in the upper-half plane $\{(b,d) \in \mathbb{R}^2 | d \geq b\}$. The set of topologically connected components describe the behavior of the function f.

3 | CLICKSTREAMS AND TOPOLOGICAL DATA ANALYSIS

We will use *connectedness* as a feature to describe customer navigation patterns. In a clickstream setting, as a customer clicks through a sequence of pages, propensity to buy waxes and wanes. The tendency to buy or not buy may be persistent or transient. We treat persistence (signal) and transience (noise) as topological features of a user's behavior within a session. To this end, we convert the session-level URL sequences into one-step transition probabilities and derive a likelihood ratio statistic to measure which is used to dissect users' sessions into signal and noise segments. The next section elaborates the derivation of the statistics to compute the persistence diagram.

3.1 | Markov chains and feature construction

As noted previously, web sessions evolve as visitors traverse the website by clicking on links successively on pages. Thus, a session is a finite sequence of interlinked pages. So, arrival at an intermediate page is conditioned on being on a prior page. Markov chains postulate that the joint probability of a sequence of pages can be decomposed into a product of conditionally independent probabilities. Therefore, a Markov chain is specified by *conditional independence* and *order*. The notion of conditional independence assumes that the sequence of clicks is not a series of independent events, and conditional independence is parameterized by degree of influence of past clicks (order). The class-conditional joint probabilities of session-level page clicks outlined in Section 2 are used to derive topological features for classification. We use transition probabilities constructed from the two classes C_1 and C_2 in the *training* set. Consider a new session denoted by U consisting of a sequence of pages $(U_1, U_2, \dots U_k)$. The joint probability $P(U_1, U_2, \dots, U_i|C_1)$ resolves as follows by the application of a first-order Markov chain:

$$\begin{aligned} P(U_1, U_2 \ \dots \ U_i) &= P(U_i|U_{i-1} \ \dots \ , U_3, U_2, U_1) \\ &\quad \times P(U_{i-1}|U_{i-2}, \ \dots \ , U_3, U_2, U_1) \ \dots \\ &\quad \times P(U_3|U_2, U_1) \times P(U_2|U_1) \times P(U_1) \\ &= P(U_1)\prod_{i=2}^{k} P(U_i|U_{i-1}). \end{aligned} \tag{1}$$

The first equality is simply the decomposition of the joint probability of the page sequence into the product of conditionally independent events. The second equality is by the property of first-order Markov chains. The class label C_1 is dropped to avoid notational clutter in Equation (1). From the session-level sequence, we build subsequences of page clicks. Subsequences $U_1, U_2, \dots U_i, 1 \le i \le k$, generate an event space $\mathbb{E}$ of subsets: $\{U_1\}$,$\{U_1, U_2\}$,$\{U_1, U_2, U_3\}$, … $\{U_1, U_2 \dots U_k\}$. The function mapping $f : \mathbb{E} \to \mathbb{R}$ generates a scalar sequence of log-likelihood ratios Λ_i given by:

$$\begin{aligned} \Lambda_i &= \log\left\{\frac{P(U_1, U_2 \ \dots \ U_i)|C_1}{P(U_1, U_2 \ \dots \ U_i)|C_2}\right\}, \quad i = 1, 2, \ \dots \ , k \\ &= \log\left\{\frac{P(U_1)\prod_{i=2}^{k} P(U_i|U_{i-1})|C_1}{P(U_1)\prod_{i=2}^{k} P(U_i|U_{i-1})|C_2}\right\}. \end{aligned} \tag{2}$$

The second expression in Equation (2) is obtained using Equation (1). Each $\Lambda_i \gtrless 0$ as a sequence evolves captures the likelihood of buying or not buying. Thus, the sequence of log-likelihood ratios $\Lambda_1, \Lambda_2, \ \dots \ , \Lambda_k$ is a summary of intrasession kinetics that serve as intermediate statistics to extract topological features.

3.2 | Sample persistence diagrams

The trajectory of a session, converted into a sequence of likelihood ratios $\{\Lambda_i\}_{i=1}^{k}$ in Equation (2), is pieced together to derive connected components from which a two-dimensional sample persistence diagram is constructed. Given a finite sequence of $\{\Lambda_i\}_{i=1}^{k}$, the sublevel sets generate birth and death pairs $(b_1, d_1), (b_2, d_2), \dots, (b_l, d_l)$. The birth and death pairs are plotted on a plane to give the persistence diagram $\mathcal{P}$ (see Figure 1). The persistence diagram $\mathcal{P}$ is a two-dimensional representation of the birth and death pairs. Each pair $b_i, d_i, i = 1, 2, \dots, l$ is plotted on a graph where the horizontal axis represents births and the vertical axis deaths. Persistence is quantified by the vertical distance from the diagonal line $b = d$ (on the right part of Figure 1). Large vertical distances (off-diagonal points) correspond to persistent segments, while small vertical distances are noise. In the context of a URL sequence within a session, a birth corresponds to a beginning of a certain behavior (ie, buy), while death is the termination of that tendency. The repeating patterns of birth and death pairs are indicative of customer engagement. In order to determine persistence segments, we calculate $100(1-\alpha)\%$ confidence intervals.[10] Any off-diagonal vertical distance in $\mathcal{P}$ that lies outside the confidence limit is tagged as a persistent signal. We distinguish persistent buying paths and nonbuying paths by color coding them by green and red dots. A persistence diagram therefore allows us to track individual customer engagement at a session level.

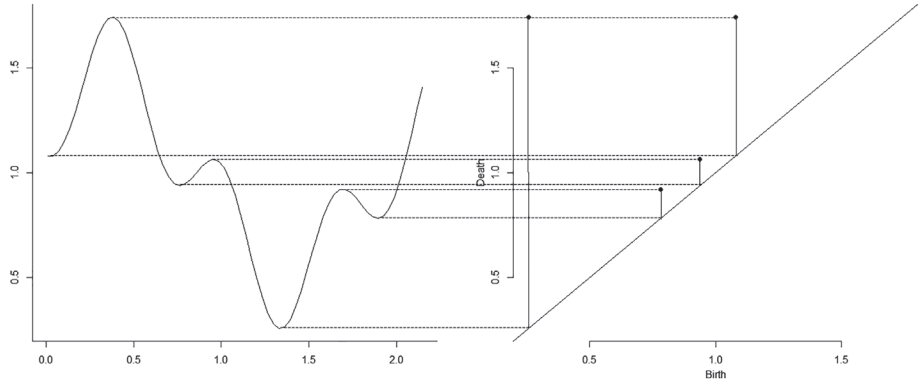

FIGURE 1 A sample persistence diagram for a buy session

Previously, applications of persistent homology[11-13] analyzed lengthy sequences of N observations. Subsamples from the long sequences are used to estimate the persistence diagram corresponding to all N points. However, in a clickstream setting, we have multiple sessions of short lengths of size $n \approx 40$ clicks per session. We therefore construct persistence diagrams from short sequences for statistical inference and prediction, unlike predecessor approaches. Persistent topological features can index segments that are temporally persistent or evanescent. The session-level persistence is banded by a $100(1-\alpha)\%$ confidence interval (CI) to determine those segments that are statistically significantly persistent. In a real-world setting, a marketer would intervene in those instances (points) in the persistence diagram outside the confidence bounds to nudge the customer back into the buying trajectory.

3.3 | Persistence diagrams and confidence intervals

At the outset, the persistence diagram $\mathcal{P}$ is a statistic estimated by $\hat{\mathcal{P}}$. Next, we compute the bottleneck distance[10] between $\mathcal{P}$ and $\hat{\mathcal{P}}$ such that it is within the interval [0,c] with probability $1-\alpha$. That is, we find a covering containing $\mathcal{P}$ with high probability. Operationally, we compute the confidence interval of the persistence diagram $\mathcal{P}$ by bootstrapping.[14] For any individual session i, we draw $\mathcal{B}$ subsamples (sessions) $S_{i1}, S_{i2}, \ldots, S_{i\mathcal{B}}$ and compute a persistence diagram $\hat{\mathcal{P}}_{ij}, j = 1, 2, \ldots, \mathcal{B}$ from each subsample. Also, we compute the persistence diagram $\hat{\mathcal{P}}_i$ of the entire session i''. The bottleneck distances $W_\infty(\bullet)$ between the persistence diagrams from subsamples and the persistence diagram from the entire input data are given by $\{W_\infty(\hat{\mathcal{P}}_{ij}, \hat{\mathcal{P}}_i), j = 1, \ldots, B\}$. Finally, the confidence interval is obtained from the $1-\alpha$ quantiles of the bottleneck distances from the $\mathcal{B}$ samples. The result is visualized by adding a confidence band around the 45° diagonal line. This is the persistence diagram! At the group level, we want to distinguish the *buyer* and *nonbuyer* groups. Toward that end, we utilize summary statistics known as *landscape* and *silhouette* introduced by Bubenik.[15]

3.4 | Landscapes and silhouettes

Both landscapes and silhouettes are real-valued functions that further summarize the information in a persistence diagram. The persistent landscape is a sequence of piecewise linear continuous functions $\lambda : \mathbb{Z}^+ \times \mathbb{R} \to \mathbb{R}$, which codify persistence diagrams. Denote the coordinates of the off-diagonal points in persistence diagram as $D = \{b_i, d_i\}, i = 1, \ldots, n$. To construct a persistent landscape, we create a set of bump functions by tenting each point $\left(\frac{b+d}{2}, \frac{d-b}{2}\right)$ for each birth and death pair b and d. The piecewise linear functions $\Lambda_p(t)$ are plotted for each value t. Before the birth and after the death point, the bump function takes value 0 and in the middle it reaches the peak.

$$\Lambda_p(t) = \begin{cases} t-b & t \in \left[b, \frac{b+d}{2}\right] \\ d-t & t \in \left(\frac{b+d}{2}, d\right] \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

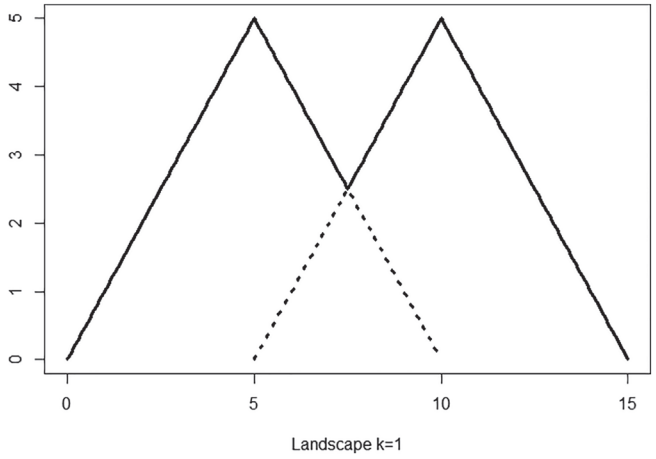

FIGURE 2 First landscape with $k = 1$ (solid line) and $k = 2$ (dotted line)

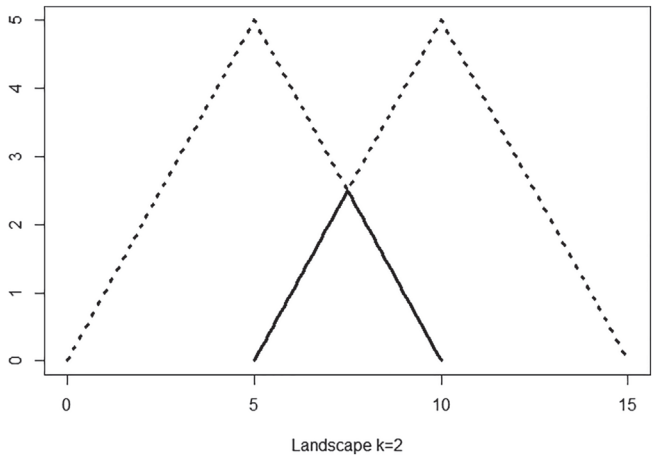

FIGURE 3 Second landscape

A summary of the collection of bump functions is the persistent landscape. The landscapes are Lipschitz continuous which reflects the robustness of persistence as the variable t is varied. For each t the kth landscape takes the kth largest value from the set of bump functions, $\Lambda_p(t)$.

$$\lambda^{(k)}(t) = k\text{th} \max_{0 \le i \le n} \{\Lambda_{p_i}(t)\}. \tag{4}$$

See Figures 2 and 3 as examples for the first and second landscape, respectively. The persistence landscape is reminiscent of *order* statistics in mathematical statistics. In clickstream data, suppose we have N sessions in total. For each session j we calculate the top two landscapes $\lambda_j^{(1)}(t)$ and $\lambda_j^{(2)}(t)$. The unknown population mean landscape $\mu^{(k)}(t) = \mathbf{E}[\lambda^{(k)}(t)], t \in [0, T]$ is calculated empirically from the sample as a *point* estimate, given by:

$$\overline{\lambda}_N^{(k)}(t) = \frac{1}{N} \sum_{0<j<N} \lambda_j^{(k)}(t). \tag{5}$$

Bubenik[15] showed the landscape $\overline{\lambda}_N(t)$ converges pointwise to $\mu(t)$ by the weak law of large numbers (WLLN). Chazal et al[11] show that $\{\sqrt{n}[\overline{\lambda}_N(t) - \mu(t)]\}_{t\in[0,T]}$ converges weakly to a Gaussian process on $[0,T]$. This justifies using the mean landscape as an estimate of the global property of a group (buyers and nonbuyers) and thus using it as a statistic to distinguish the groups.

The *training* sets of buyer and nonbuyer sessions are then used to construct an average persistence landscape for each class $C_i, i = 1, 2$. For a given session i in the *testing* set, the distances of its persistence landscape statistic λ_i to the average persistence landscape statistics $\overline{\lambda}_i, i = 1, 2$ of the two classes C_1, C_2 are measured. The session is classified into that class for which the distance is smaller.

Another summary function, called silhouette, is a weighted average summary function over all bump functions.[11] For a diagram with off-diagonal points $p_i, i = 1, \dots, n$, where $p_i = \{b_i, d_i\}$, the silhouette is defined as

$$\phi(t) = \frac{\sum_{i=1}^{n} |d_i - b_i|^p \Lambda_{p_i}(t)}{\sum_{i=1}^{n} |d_i - b_i|^p}. \tag{6}$$

The choice of the parameter p can be thought of as a trade-off between weighting all the segments formed by the birth and death pairs equally and weighting them unequally. For small values of p, $\phi(t)$ is dominated by the effect of low-persistent (transient) features, while large values of p emphasize the high-persistent features. Note that as p increases, the silhouette approaches the first landscape $\lambda^{(1)}(t)$, that is, $k = 1$. The silhouette also satisfies Lipschitz continuity, thus maintaining asymptotic properties due to WLLN. As we will see in the following section, landscapes and silhouettes are useful tools for separating buyers and nonbuyers and lend themselves to classification.

3.5 | Persistence diagram results

At an individual level, we track a customer's clicks as the session progresses. Within each session, we calculate $\Lambda_i, i = 1, 2, \dots, k$ and it forms a scalar-valued function $f : \mathbb{R} \to \mathbb{R}$. The connectivity of the sublevel sets $f^{-1}(-\infty, a]$ as the parameter a varies between 0 to ∞ reflects changes in buying intentions.

The evolution of a session is captured by the persistence diagram. We distinguish the buying and nonbuying patterns in the following manner. If a local maximum occurs before a local minimum, it is denoted by a red dot indicating a tendency not to buy. Similarly, if a local minimum is followed by a local maximum, it is denoted by a green dot, indicating a propensity to buy. The visually discernible buying and nonbuying patterns and their persistence are simply the vertical distances from the 45° diagonal line. Figures 4 and 5 are examples of typical nonbuying and buying sessions. The horizontal axis is the number of pages (clicks) and the vertical axis is the quantity $\Lambda_i, i = 1, 2, \dots, k$ (with the ratio of buy to no-buy conditional on the number of pages clicked). This is clearly a discrete version of the function (f) from which the persistence diagram is derived. Corresponding to the no-buy session depicted in Figure 4, we display the persistence diagrams in a series of four panels in Figure 6. In the last three panels, top right, bottom left, and bottom right, we notice a green dot persisting outside the confidence band. This green dot corresponds to click 7 in Figure 4. Notice the ratio statistic Λ_i achieves its maximum at click number 7 where birth and death values are 1.52 and 12.37 respectively. Similarly, a red dot appears in the bottom left and bottom right panels. The onset of the red dot (bottom left) starts at the edge of the confidence band where birth and death values are respectively 6.25 and 12.37. This corresponds to click 9 in Figure 4. In the bottom right panel, the coordinates of the red dot (3.16 and 12.37) are associated with click number 11 in Figure 5. The takeaway from this analysis is that the customer was well on his way to buying until click 7. However after a session depth of seven clicks, interest presumably diminished and the session turned into a no-buy session. Appropriate intervention (discount coupon, free shipping) in the immediate aftermath of click 7 may have converted the customer. The session that resulted in a buy (Figure 5) may be interpreted similarly (Figure 7).

At the group level, we want to distinguish the buyer group and the nonbuyer group by persistence homology. We use the persistence landscape and silhouette for topological feature extraction. Both statistics based on topological

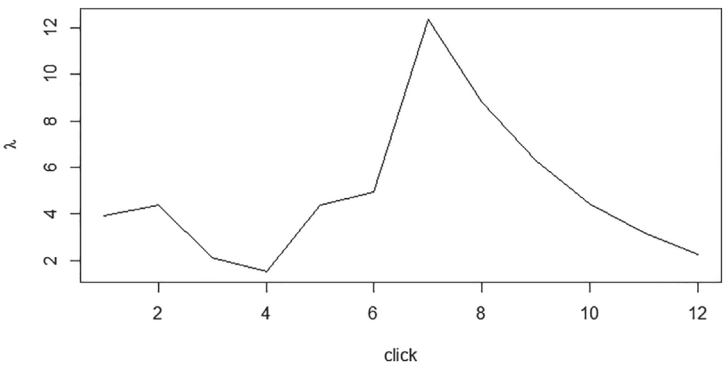

FIGURE 4 Example of a nonbuying session

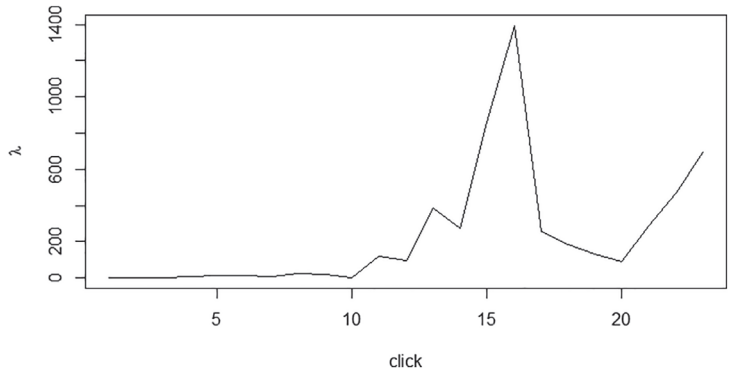

FIGURE 5 Example of a buying session

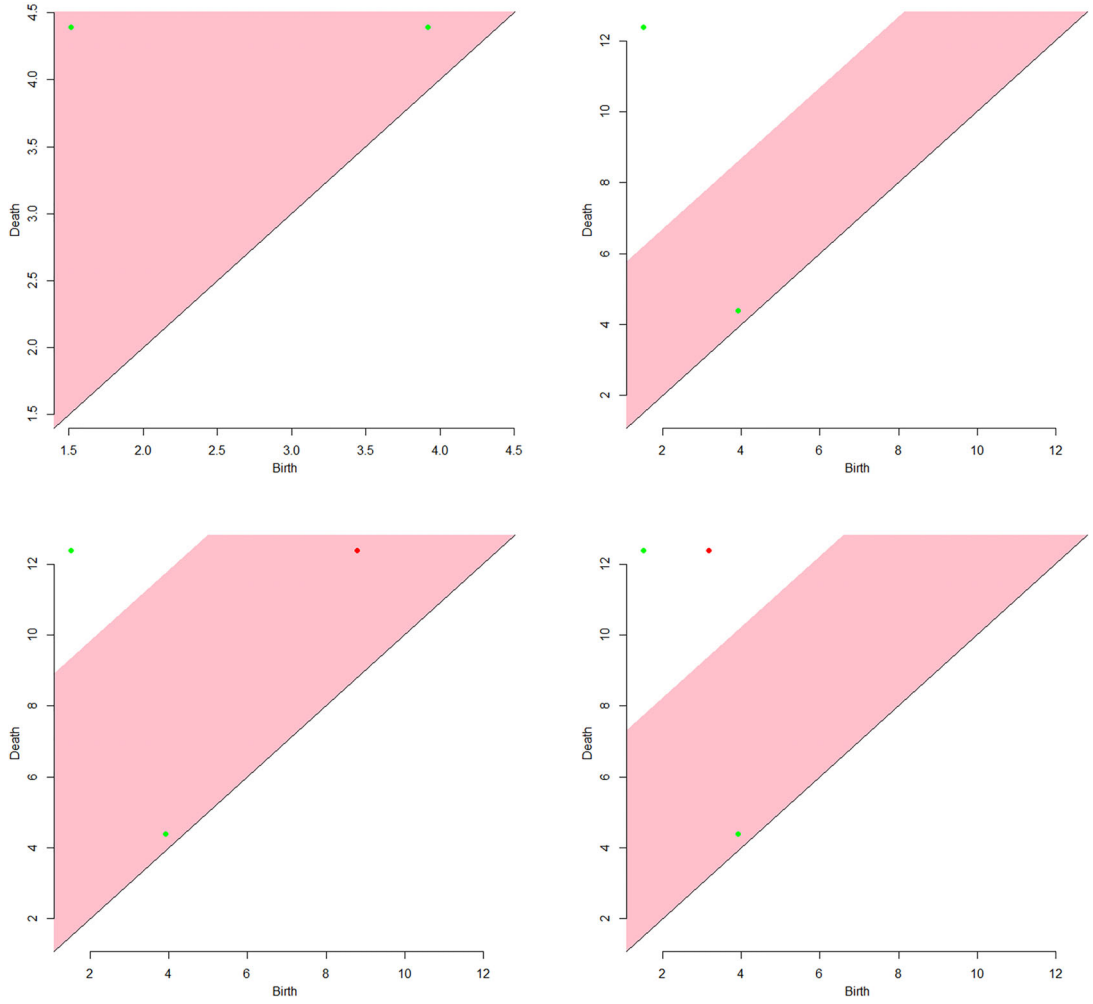

FIGURE 6 Example persistence diagram corresponding to a nonbuying session [Color figure can be viewed at wileyonlinelibrary.com]

summaries attempt to measure group separation. For the buy and no-buy groups, we calculated the mean landscape with the parameter value k equal to 1 and 2 in Equation (5). Note that in Equation (3) the silhouette statistic depends on the parameter p. We set p equal to 0.001 and 1, respectively. These settings gave the best distinction between the buyers and nonbuyers. From the Figure 8, the visual differences are evident among average landscapes and silhouettes at different parameter settings. The landscape with $k=1$ considers only the most persistent pairs in the persistence diagram. It has a single peak which reflects that the dominant intention of customers in the two groups is unique and different. The landscape with $k=2$ treats all pairs in the persistence diagram uniformly

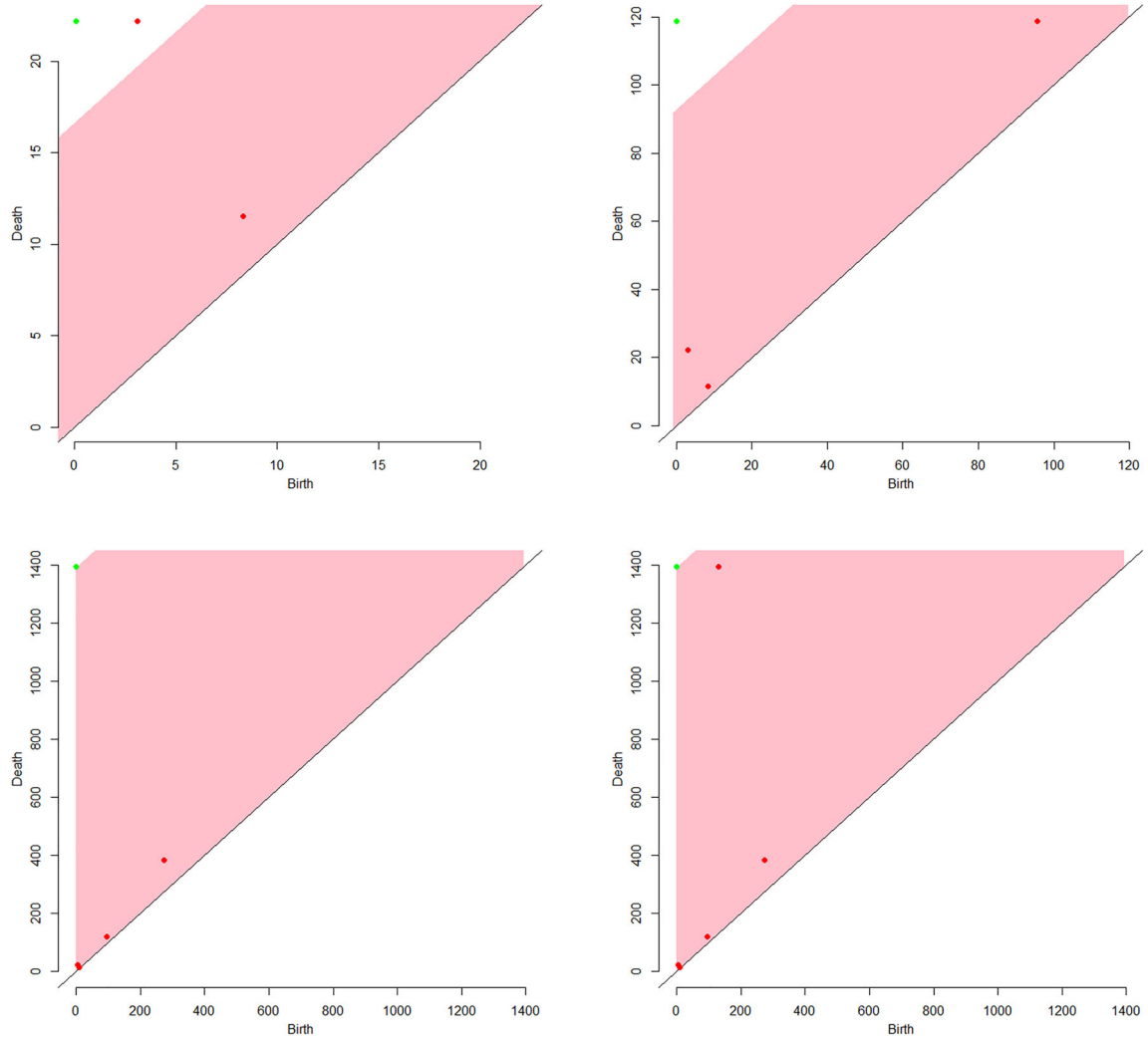

FIGURE 7 Example persistence diagram corresponding to a buying session [Color figure can be viewed at wileyonlinelibrary.com]

and captures subtle changes. The multiple peaks in Figure 8 may correspond to specific behavioral patterns in the *buy* and *no-buy* groups. The silhouette with a small value of p is a weighted average of landscapes with different values of the parameter k. Note that as the parameter p gets large, the silhouette converges to the first landscape with k equal to 1.

As we use the mean landscape to distinguish groups of buyers and nonbuyers, we invoke statistical inference. To test the null hypothesis that the average landscapes are resulting from identical distributions, we use permutation tests. The permutation principle states that if two group averages are generated from identical distributions then the observed data should arise from the same distribution with high probability across n shufflings (permutations) of the data (sessions). To test the hypothesis, we randomly sample the URL sequences of individual sessions from groups of buyers and nonbuyers n times. For each sample we compute the difference (L_2) distance between average landscapes ($k = 1$). At 2000 permutations, the landscapes were different 95% of the times confirming that they come from different distributions. Having determined the usefulness of mean landscapes to detect differences between buyers and nonbuyers, we then utilize them for classification. The average persistence landscape and silhouette are used as the main tools to classify sessions. We randomly choose 1000 sessions from both groups and calculate the mean landscape and the mean silhouette. Then, we choose another 1000 random sessions as the test set and classify them according to the L_2 distances from their landscape or silhouette to the group average. The results are averaged over 10 experiments and summarized into recall $\left(\frac{\text{TP}}{\text{TP+FP}}\right)$, precision $\left(\frac{\text{TP}}{\text{TP+FP}}\right)$, FPR $\left(\frac{\text{FP}}{\text{FP+TP}}\right)$, and F-measure (the harmonic mean of recall and precision). The results, seen in Figure 9, show high accuracy and stability.

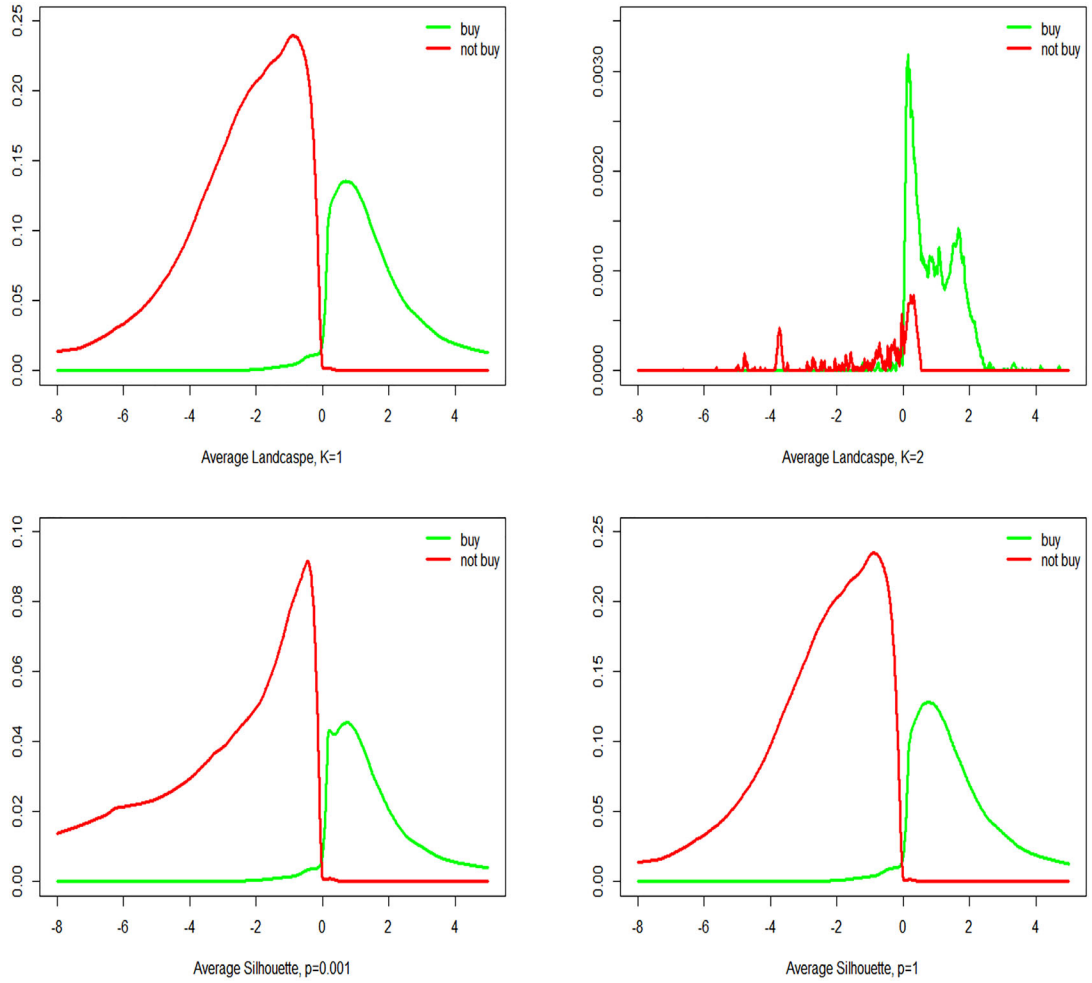

FIGURE 8 Average landscapes and silhouettes [Color figure can be viewed at wileyonlinelibrary.com]

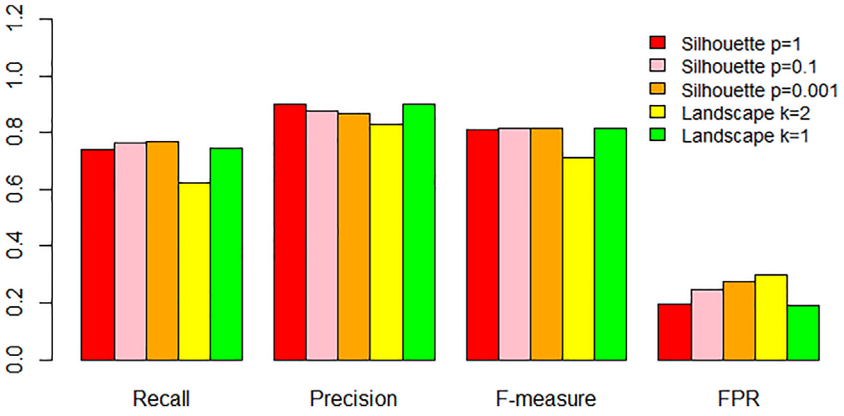

FIGURE 9 Performance of classification model [Color figure can be viewed at wileyonlinelibrary.com]

3.6 | Markov chain models

We invoke Markov chains to predict the likelihood of conversion of a visitor in a given session. Treating a website as a graph where the URLs are *nodes* and *edges* are transition probabilities, the simplest Markov model is of the first order in which the probability of being on the current URL page depends only on the previous page clicked (page transition probability). Higher order chains trace history to previously clicked pages 2, 3, … , n. Furthermore, as any website is not a fully connected graph, transition probability matrices are sparse. So, we use the Chapman-Kolmogorov equations[16] as a

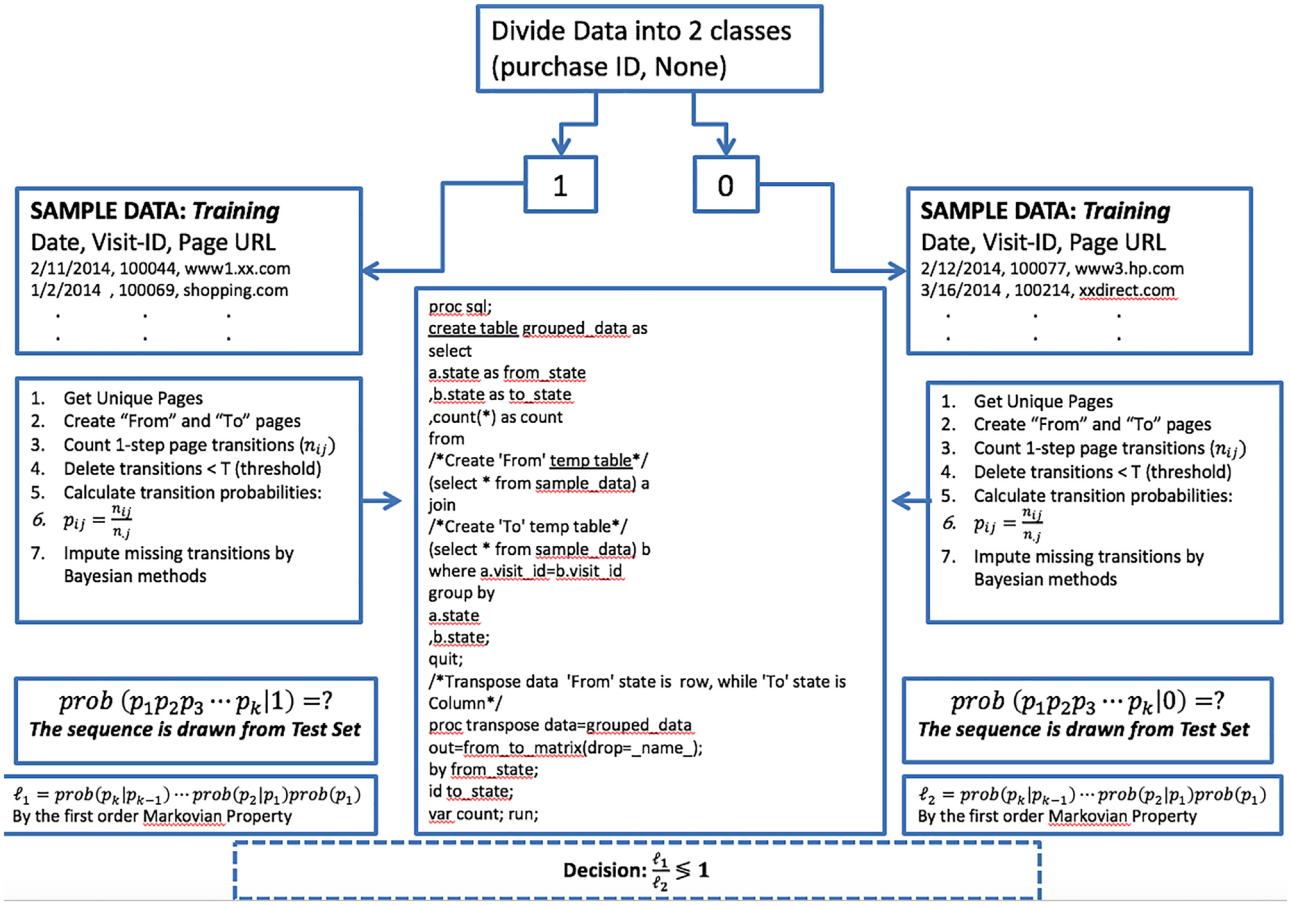

FIGURE 10 A flowchart describing the proposed classifier [Color figure can be viewed at wileyonlinelibrary.com]

tool for missing value imputation to enrich graph connectedness. Let a new session denoted by U consist of the sequence of pages $(U_1, U_2, \ldots U_k)$. We compute the class-conditional joint probabilities

$$P(U_1, U_2, \ldots, U_k|C_1), \quad P(U_1, U_2, \ldots, U_k|C_2)$$

denoted by L_1, L_2, respectively. The decomposition of joint probability is given by Equation (1). We drop the class label C_1 for simplicity of notation. Figure 10 is an overview of the modeling methodology including the decision rule. The integers 1 and 0 in the flow chart represent the *Buy* and *NoBuy* classes. The decision rule for classifying sequences is determined by the ratio $D = \frac{L_1}{L_2}$. If $D \leq 1$ then $U \in C_2$, else $U \in C_1$.

3.7 | Markov chain results

The product category of supplies has a conversion rate of ~1%. We report the quality metrics recall, FPR, and F2-measure. For the purposes of scoring sessions in the *testing* set, we divide them into lengths of 5, 10, 15, 20, 25, 30, 35, 40, and 45 pages to evaluate likelihoods at different depths of an active session. At each session length, conversion likelihoods (see Figure 10 for an outline of classification methodology) are calculated. We limited the summaries to session depth up to 45 as there was no empirical advantage to analyzing sessions beyond the 45-page limit. Also, we only report results for Markov chains up to three *orders* since performance decays at higher *orders*. More specifically, for the product category of supplies with the application of fourth-order Markov chains, recall is in the (0.55, 0.66) range, while FPR is dismally high in the range of (0.50, 0.66) over the session depths of 5 to 35.

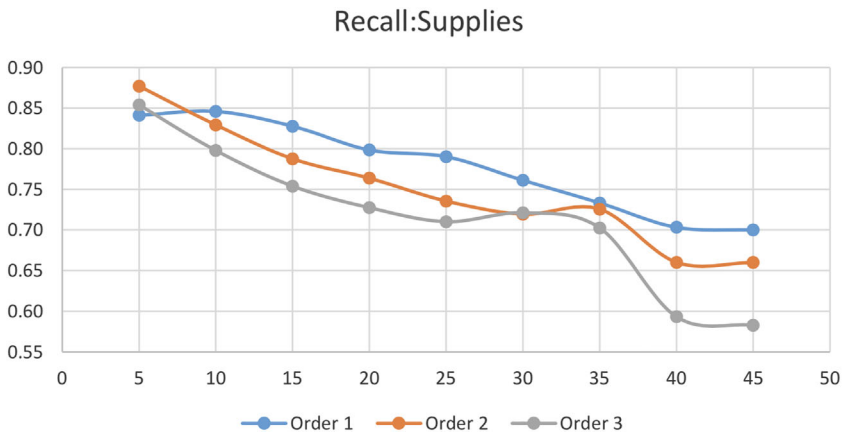

FIGURE 11 Recall: performance is ~ equal across first, second, and third orders [Color figure can be viewed at wileyonlinelibrary.com]

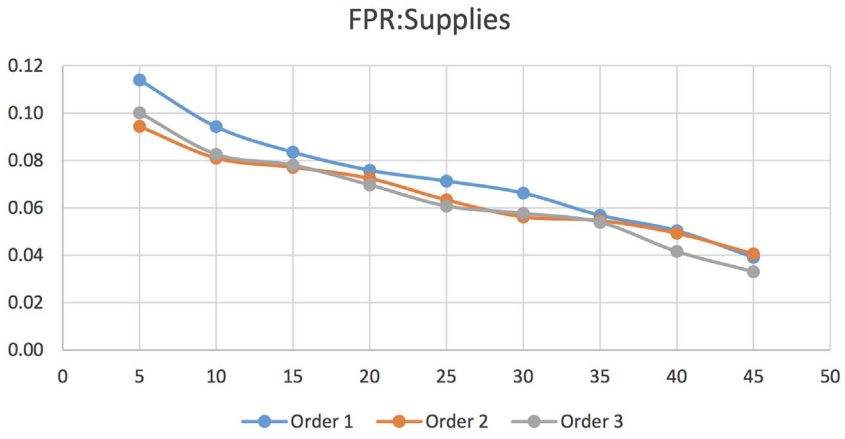

FIGURE 12 FPR: first-order MC is within range of higher order chains [Color figure can be viewed at wileyonlinelibrary.com]

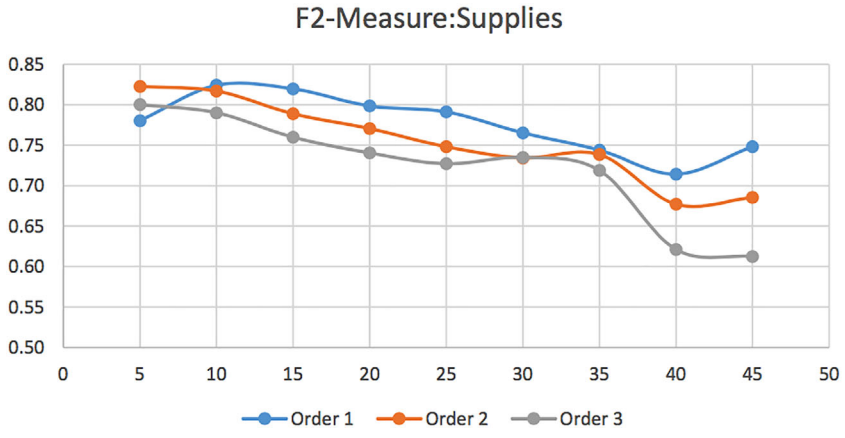

FIGURE 13 F2-measure: first-order MC delivers superior performance [Color figure can be viewed at wileyonlinelibrary.com]

Figures 11,12, and 13 capture the metrics as a function of the session depth, with the *x*-axis representing the partial length of sequence within a session. The metrics presented are averages over several runs. Each run consists of *training* and *testing* sets resampled (simple random sampling with replacement). The resampling is to analyze the stability of the classifier in terms of the standard error (SE). We report averaged results from a set of six runs (additional runs did not yield significant differences). The recall for the first-order chain is higher than the chain of second and third orders. The higher order chains track each other closely, but produce lower recall. The FPR for the first-order chain is higher but within a small delta from chains of higher orders. The F2-measure is quite satisfactory. It may be observed that the likelihood of conversion is highest in the 5 to 35 page range. Beyond session length of 35 pages, the metrics stabilize. The insight here is that customers 5 to 35 pages deep in a session are prime targets for marketing intervention. There is no additional information gained by tracking sessions beyond a depth of 40 pages. More data do not mean more information!

Finally, the standard errors are in the range [0.001, 0.04] and [0.01-0.02], respectively, for recall and FPR: we report the ranges on the standard errors as they are different at each session length across multiple runs. In the numbers reported we pick the range between the lowest and highest standard errors. We noticed that as session length increases, the standard error of the corresponding metric increases. The standard errors are only reported for chains of order one. Similar variability was observed for the higher order chains. It may be concluded that despite the low conversion rate in the supplies category, Markov chains deliver good performance.

3.8 | Imputing missing page transitions

Commercial websites are living and breathing entities that change constantly as old pages are updated with new links to establish new connections. When two pages U_i and U_j are linked by intermediate page(s) $U_k \in \mathbb{S}$, then page U_i may be updated with a link pointing to newly created page U_k. This phenomenon has consequences in modeling. Pages that were absent in the *training* set (historical data) are likely to appear later because some pages were updated to include links connected to the new ones. For example, a transition from page $U_i \to U_j$ given by p_{ij} may be missing in the *training* set, but appears in the scoring phase. This situation is a vexing issue in practice. A routinely used quick-fix is to impute the missing transition probability by an arbitrary small number such as 10^{-6}. A more methodical approach is to invoke the Chapman-Kolmogorov (C-K) equations.[16] C-K equations are a tool to estimate (n-level, $n \geq 2$) transition probabilities. If pages $i \to k$ are linked and pages $k \to j$ are linked, C-K imputes p_{ij} from $i \to k \to j$ by marginalizing over all $k \in \mathbb{S}$, $\mathbb{S}$ is the collection of all pages on the website. In the following we will briefly recall computing nth-order probabilities by Chapman-Kolmogorov equations. For any $n \geq 0$, $m \geq 0$, $i \in S, j \in S$,

$$\begin{aligned} P_{ij}^{n+m} &= P(X_{n+m} = j | X_0 = i) \\ &= \sum_{k \in \mathbb{S}} P(X_{n+m} = j, X_n = k | X_0 = i) \\ &= \sum_{k \in S} P(X_{n+m} = j | X_n = k, X_0 = i) P(X_n = k | X_0 = i) \\ &= \sum_{k \in S} P_{kj}^m P_{ik}^n. \end{aligned}$$

In the context of clickstreams, if a direct link between pages U_i *and* U_j is missing, we will capture all the two-level paths $U_i \to U_k \to U_j \; \forall k \in \mathbb{S}$. We demonstrate the imputation results for the product group of supplies. The analysis is conducted using two methods of imputation: first, the standard approach of imputing missing transition probabilities by a small value, 10^{-6} in our studies, and second, the Chapman-Kolmogorov equations. Recall that application of the C-K equations is relevant only for first-order Markov chains. We report only the overall performance metric (F2-measure), seen in Figure 14. For the product category of supplies, there is a noticeable lift in the F2-measure in the session-length range between 5 and 35 pages (see Figure 13 for a comparison). The other metrics and product categories are excluded as we observed similar trends across the remaining product lines. The results show that there is an improvement in performance due to imputation by the Chapman-Kolmogorov equations. Although we see a lift in performance, it is important to point out that imputation may not necessarily yield better classification performance. It is expected to deliver a more reliable estimate of a probability of a missing transition that may appear in the scoring phase. In summary, the Chapman-Kolmogorov equations appear to be a useful tool for missing value imputation.

4 | COMPARISON WITH EXISTING METHODS

In this section, we compare topological data analysis with Markov chains. As mentioned earlier, we evaluated the performance of the classifiers as sessions progress, that is, breaking sessions into lengths of $5, 10, 15, 20, \ldots$ and applying topology-based and Markov chain classifiers. We compare Markov chains of *order* one with average topological landscapes ($k = 1$), since we saw that first-*order* Markov models deliver best performance.

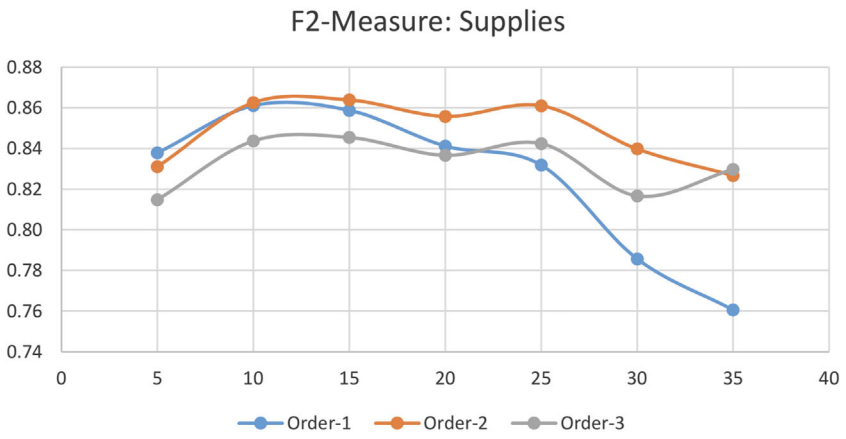

FIGURE 14 F2-measure for supplies after imputation, showing clear improvement [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Combined session results

Method	Landscapes				Markov chains			
Page number	Recall	Precision	FPR	F-measure	Recall	Precision	FPR	F-measure
5	0.48	0.80	0.22	0.60	0.88	0.27	0.09	0.42
10	0.74	0.93	0.11	0.82	0.81	0.4	0.08	0.53
15	0.87	0.96	0.06	0.91	0.75	0.45	0.07	0.56
20	0.91	0.98	0.04	0.94	0.7	0.47	0.07	0.56

Table 2 shows performance in terms of recall, precision, FPR, and F-measure. While Markov chains beat landscapes relative to Recall and FPR for short sessions, the trend is opposite for precision and F-measure. This suggests that while landscapes are better at reducing false positives, they fail to detect true buyers. This is likely because buyers and nonbuyers behave similarly in the early stages of a session, but the situation changes as session lengths increase. More specifically, building landscapes to capture topological signatures requires longer sessions to compute birth and death patterns (local maxima and minima), while Markov chains are based purely on one-step transition probabilities. Therefore, prediction at the beginning of a session (5 to 10 page clicks) by average landscapes is worse, but as the session builds, they allow us to extract persistent patterns and outperforms Markov chains. The results confirm that topological landscapes beat Markov chains starting at session lengths 15 and beyond, and are better suited to capture long-term behaviors in sessions.

In conclusion, our article proposes practical methods to deal with real-world data obtained from a large enterprise website. The adoption of TDA to clickstreams produced surprisingly good results. Extending the one-step transitions probabilities to construct topological functionals to tease out oscillations in navigation patterns adds to data scientists' arsenal of tools and opens areas of research for further investigations. More importantly, it helps marketing ROI. The entire pipeline from data collection to preprocessing to modeling and implementation was an imposing exercise. Hopefully, our learnings, recipes, and methodology will serve as a guide to practical implementation in commercial settings. As we could not cover several aspects, detailed reports, experimental results, and code, they will be made available upon request, including the data for testing the reproducibility of the results we presented.

ORCID

Choudur Lakshminarayan https://orcid.org/0000-0003-2571-9372

REFERENCES

1. Montgomery AL, Li S, Srinivasan K, Liechty JC. Modeling online browsing and path analysis using clickstream data. *Mark Sci*. 2004;23(4):579-595. https://doi.org/10.1287/mksc.1040.0073.
2. Srivastava J, Cooley R, Deshpande M, Tan PN. Web usage mining: discovery and applications of usage patterns from web data. *ACM SIGKDD Explor Newslett*. 2000;1(2):12. http://portal.acm.org/citation.cfm?doid=846183.846188.
3. Moe WW, Fader PS. Dynamic conversion behavior at E-commerce sites. *Manag Sci*. 2004;50(3):326-335. https://doi.org/10.1287/mnsc.1040.0153.

4. Borges J, Levene M. Evaluating variable-length Markov Chain models for analysis of user web navigation sessions. *IEEE Trans Knowl Data Eng*. 2007;19(4):441-452. http://ieeexplore.ieee.org/document/4118703/.
5. Brusilovsky P, Kobsa A, Nejdl W. The adaptive web: methods and strategies of web personalization. No. 4321 in Lecture notes in computer science. *State-of-the-Art Survey*. Berlin, Germany; New York, NY: Springer; 2007 OCLC: ocn124038344.
6. Deshpande M, Karypis G. Selective Markov models for predicting wb page accesses. *ACM Trans Internet Technol*. 2004;4(2):163-184. http://portal.acm.org/citation.cfm?doid=990301.990304.
7. Eirinaki M, Vazirgiannis M, Kapogiannis D. Web path recommendations based on page ranking and Markov models. Paper presented at: Proceedings of the 7th ACM International Workshop on Web Information and Data Management – WIDM '05; 2005:2; Bremen, Germany, ACM Press. http://portal.acm.org/citation.cfm?doid=1097047.1097050.
8. Lakshminarayan C, Kosuru R, Hsu M. Modeling complex clickstream data by stochastic models: theory and methods. Paper presented at: Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion Montréal; 2016:879-884; Québec, Canada, ACM Press. http://dl.acm.org/citation.cfm?doid=2872518.2891070.
9. Edelsbrunner H, Harer J. Persistent homology–a survey. In: Goodman JE, Pach J, Pollack R, eds. *Contemporary Mathematics*. Vol 453. Providence, Rhode Island: American Mathematical Society; 2008:257-282. http://www.ams.org/conm/453/.
10. Fasy BT, Kim J, Lecci F, Maria C. Introduction to the R package TDA; January 2015. arXiv: 1411.1830. http://arxiv.org/abs/1411.1830.
11. Chazal F, Fasy BT, Lecci F, Rinaldo A, Wasserman L. Stochastic convergence of persistence landscapes and silhouettes. Paper presented at: Proceedings of the Annual Symposium on Computational Geometry - SOCG'14; 2014:474-483; Kyoto, Japan, ACM Press. http://dl.acm.org/citation.cfm?doid=2582112.2582128.
12. Wang Y, Ombao O, Chung MK. Persistence landscape of functional signal and its application to epileptic electroencaphalogram data. Student paper; 2013.
13. Kovacev-Nikolic V, Bubenik P, Nikolić D, Heo G. Using persistent homology and dynamical distances to analyze protein binding. *Stat Appl Genet Molecul Biol*. 2016;15(1):19-38. https://doi.org/10.1515/sagmb-2015-0057.
14. Efron B, Tibshirani R. *An Introduction to the Bootstrap. No. 57 in Monographs on Statistics and Applied Probability*. New York, NY: Chapman & Hall; 1993.
15. Bubenik P. Statistical topological data analysis using persistence landscapes; 2015. arXiv: 1207.6437. http://arxiv.org/abs/1207.6437.
16. Ross SM. *Introduction to Probability Models*. 10th ed. Amsterdam,Netherlands; Boston, MA: Academic Press; 2010 OCLC: ocn444116127.

How to cite this article: Lakshminarayan C, Yin M. Topological data analysis in digital marketing. *Appl Stochastic Models Bus Ind*. 2020;1–15. https://doi.org/10.1002/asmb.2563