A Novel Neural Network Model for Customer Base Analysis and User Profiling

Paramveer Dhillon and Sinan Aral
{dhillon|sinan}@mit.edu
MIT Sloan School of Management

Job Market Paper[1]
June 20, 2018, Draft Version: 1.0

[1]The authors would like to thank Michael Zhao, Dean Eckles, Peter Fader, and John Hauser for constructive feedback. All errors remain our own.

Abstract

The digital breadcrumbs left by various online activities have provided an unprecedented opportunity for marketing managers to understand the consumption patterns of their customer base. This paper proposes a novel neural network based modeling framework that combines customers' time-invariant attributes with their time-varying consumption activity to model the dynamics of their engagement patterns. Our framework combines neural networks with a dictionary learning procedure that generates interpretable time-varying distributions over latent customer interests—avoiding the non-interpretability neural networks are generally criticized about. The proposed model is both computationally and statistically efficient: it is fast to estimate, scales to large data sizes, and can harness external data sources as an empirical prior. These advantages make our method well-suited to the challenges posed by modern datasets. We highlight the utility of our approach by modeling the content consumption patterns of Boston Globe website readers. Results show that our framework unravels subtle trends in customer behaviors and estimates interpretable dynamic customer profiles. Further, our approach provides a statistically and economically significant improvement in predicting customers' future visitation and churn prospects compared with other state-of-the-art methods.

Keywords: *Customer Analytics; Consumer Preferences; Digital Marketing; Online Content; Machine Learning; Deep Learning; Natural Language Processing.*

1 Introduction

The advent of the Internet and digitization of consumer activity has provided a golden opportunity for companies to gather more information about customers. Digital platforms can use the abundant click-by-click data collected from consumers for a variety of purposes. For instance, they can track how consumers interact with their website and accordingly make adjustments to improve the user experience to maintain a sustained level of user engagement. They can also use consumer data to make product recommendations (Bodapati, 2008), assess the churn probability and hence customer lifetime value (Moe, 2003; Moe and Fader, 2004), generate dynamic personalizations (Hauser et al., 2009; Urban et al., 2013), offer customizations (Ansari and Mela, 2003), target prices (Dubé and Misra, 2017), target advertisements (Goldfarb and Tucker, 2011; Perlich et al., 2014), and personalize search results (Yoganarasimhan, 2016). Beyond just its business value (Trusov et al., 2016; Martens et al., 2016), consumer data can also be leveraged for public policy ends. The digital trails left by consumers on social media websites like Twitter can be used to gain insights into their psychological and physical well-being (Schwartz et al., 2013; Sinnenberg et al., 2017).

Consumer information is increasingly viewed as an important strategic asset for companies. However, despite the exponential growth in data generation and collection over the past decade, generating actionable insights from this data faces several challenges. First, the large data sizes pose computational challenges. This is especially true for Bayesian methods where posterior inference is typically done via Markov-chain Monte Carlo (MCMC) style methods, which can have very slow convergence times. Second, online clickstreams and other user generated content (UGC) often contains significant textual information which lives in very sparse and high dimensional spaces[2]. This makes statistical inference using traditional methods hard since such methods typically estimate a parameter for each dimension. Third, the dynamic (or temporal) nature of this data further aggravates the challenges posed by the size and sparsity of data. However, it is this change in customer preferences indicated by the dynamics of content consumption that is commercially very valuable to model since it may indicate a purchase intent.

In spite of these challenges, companies have certainly managed to unlock some of the enormous potential of consumer data. Yet, it is clear that much still remains untapped. Customers' content consumption history is a mixture of their interests over a period of

[2]The standard way of encoding a *word* is via one-hot-encoding, that is, a sparse vector of size the vocabulary of English ($\sim 300K$) with all 0s, except a 1 at the location of the lexicographically sorted index of that word.

time, so it is pivotal to devise models that are able to separate these distinct interests while still being interpretable and efficient to estimate.

In this paper, we propose a novel neural network based modeling framework that addresses above shortcomings. Our framework consists of two closely related neural network architectures for the problems of modeling customers' consumption dynamics and for predicting their future engagement metrics such as churn and frequency of return visits. Our framework leverages both time invariant customer attributes such as age, gender, occupation, and the time-varying consumption history to estimate an evolving mixture distribution over customers' latent interests. The dynamic customer profiles estimated by our approach provide interpretable summaries of customers' content consumption behavior. These profiles are also predictive of customers' future engagement patterns, so we further use them to predict customers' churn and return visit propensities.

Our approach is efficient to estimate and easily scales to large data sizes as it does not involve costly sampling procedures for model inference. It addresses the data sparsity issue by embedding the high dimensional clickstream data into low-dimensional user-specific projections. It handles dynamics efficiently by borrowing statistical strength from the same customer longitudinally over time. Our approach also avoids the general standard criticism of neural networks—the lack of interpretability and the inherent "black-box" nature of their predictions—by using a novel dictionary learning procedure that enables us to estimate interpretable time-varying distributions over customers' interests. Hence, it addresses the issues posed by *size*, *sparsity*, and *dynamics* of large datasets discussed earlier and further draws on the best properties of both the neural network and probabilistic modeling paradigms.

Our work contributes to several strands of literature. First, our work contributes to the research on marketing analytics and user profiling from online clickstream data (Trusov et al., 2016; Farias and Li, 2016). Trusov et al. (2016) examine a similar research question as us but with a different modeling approach. They extend Correlated Topic Models (CTM)[3] to incorporate visitation intensity, heterogeneity, and dynamics to generate user profiles from online browsing data. However, unlike this paper, they do not explicitly model churn and return visit probabilities. Further, in terms of methodology their approach relies on Markov Chain Monte Carlo (MCMC) sampling for model inference, which is very slow to estimate and the results are highly sensitive to parameter initialization. Farias and Li (2016), on the other hand, propose a fast and efficient novel matrix

[3]A variant of Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

factorization for learning user preferences from online activity trails. However, they do not model the dynamics of these preferences or their impact on customer engagement metrics such as churn.

Next, our work contributes to the literature on customer base analysis. Much of the work in this area involves building stochastic models of customer purchase behavior using summary statistics e.g. recency, frequency, and the monetary value of the purchases (Schmittlein et al., 1987; Abe, 2009; Fader et al., 2010). These papers project the future spending of customers using only these customer-level summary statistics and typically do not model the content consumed or other digital trails left by the customers. Our approach is different from this line of work as it can flexibly (in a non-linear fashion) incorporate covariates corresponding to customers' consumption activity while still being efficient to estimate.

Our work also contributes to the literature on modeling evolution of consumers' preferences and their sensitivities to various marketing variables. The most classic work in this area is Guadagni and Little (1983), which models evolution of brand preferences using exponential smooths of customer-level brand-loyalty parameters. Since then, there has been much follow-up work on modeling evolution of brand preferences. This line of work typically assumes some explicit parametric form for preference evolution and further assumes that each customer's preference trajectory deviates from the population-level trajectory by an individual-specific offset (Neelamegham and Chintagunta, 2004; Kim et al., 2005; Lachaab et al., 2006; Sriram and Kalwani, 2007). Methodologically, these papers build a hierarchical Bayesian or a Bayesian state-space model to flexibly incorporate different kinds of heterogeneity and the time dynamics. More recently, Dew et al. (2017) have used Gaussian processes to model dynamics of consumer preferences. Though this body of work is methodologically elegant and flexibly models consumer heterogeneity—a key construct in marketing—these approaches are highly computationally inefficient. Further, unlike this paper, these approaches do not predict customer churn.

Finally, our work contributes to the burgeoning literature in marketing on using machine learning methods for studying customer preferences using various forms of user-generated content (UGC) e.g. consumer reviews, online chats or searches (Netzer et al., 2012; Tirunillai and Tellis, 2014; Büschken and Allenby, 2016; Liu and Toubia, 2017; Timoshenko and Hauser, 2017). Relatedly, this paper also contributes to the growing literature in machine learning and deep learning where neural networks have given state-of-the-art performance in several domains including natural language processing (Bengio et al.,

2003; Mikolov et al., 2013a), social networks (Grover and Leskovec, 2016), and information retrieval (Le and Mikolov, 2014). Our work is the first to the best of our knowledge that has used neural networks for modeling evolving customer preferences and more broadly used them for customer base analysis.

The rest of the paper is organized as follows. Next, we give an overview of the empirical setup of our problem and describe the data. Section 3 provides a brief introduction to neural network modeling which serves as a foundation for our model specifications described in Section 4. In Section 5 we describe the results of our model estimation on content consumption data from Boston Globe. We discuss managerial implications and provide avenues for future research in Section 6.

2 Empirical Setting

We model the dynamics of content consumption in the context of online news. Online news consumption is a perfect testbed for studying evolution of customer interests as a broad and representative base of internet users consume content online. Further, news consumption patterns do often change saliently over a period of time. For instance, there has been a substantial increase in interest in political news after the 2016 US Presidential election. Similarly, there is an up-tick in the consumption of news articles related to basketball or football right before the start of the sporting season or during playoffs. There are several reasons for these changes in news consumption patterns. They can change owing to customers' innate individual-level idiosyncrasies, for example via self-discovery or learning about a new topic on the Internet. They can also change due to broad population-level trends or they can change due to the variation in availability of certain kinds of content in certain time periods. For instance, there is usually an increased supply of football content during the playoffs and a similar glut of political news during and after a major election.

It is important to understand this ebb-and-flow of news consumption from a substantive point of view as it can unravel broader shifts in public opinion which could have implications for the health of a democracy. It is also pivotal to track these news readership dynamics from the perspective of the newspaper as it can allow them to optimize content placement on their website to increase reader engagement. It can also allow them to generate digital profiles of their customers. Such profiles may be used for targeting of advertisements or for pro-actively acting to lower the customers' churn propensity—an

application that is a focus of this paper. In this paper, we use the news consumed by readers longitudinally to create dynamic customer profiles to understand the evolution of their tastes as well as to predict their future engagement metrics.

This paper models changes in customers' consumption behavior in online news. In many other contexts such as retailing or supermarket purchases, this is often equivalent, or at least assumed equivalent to modeling changes in demand-side consumer preferences. However, in our context, a strong supply-side mechanism exists that is completely consistent with observed behavior. It is represented by external factors that change the supply of various kinds of news stories. In reality, online news consumption behavior is probably driven by both supply-side and demand-side factors. We do not model the supply-side in this paper and instead take the news content as exogenously determined each time period. The question of disentangling supply and demand is interesting, but it is beyond the scope of this paper. Modeling customer behavior itself is sufficient for a number of predictive customer analytics applications like the ones that we are interested in.

2.1 Data

We use more than 3 years worth of individual-level clickstream data from Boston Globe[4] from February 1, 2014 to May 31, 2017 to perform our analysis. Our data contains fine-grained information about the online reading behaviors of a total of unique 288,629,919 visitors (who made a total of 479,275,430 visits) over this 41 month period. This includes information regarding which articles they read, how much time they spent reading those articles, and their subscription status. We further have access to granular demographic data for the visitors such as area code, zip-code, device type (mobile or desktop), operating system, and country.

One might not expect to see interesting dynamics in content consumption preferences on a day-to-day basis since news stories typically last for a few days. It is also typical for users' interests to crystallize over time spans longer than a day. Further, some people only read the news on weekends, so we perform our analyses at the level of a week. However, we do zoom into day-level dynamics whenever necessary. We further restrict our dataset by weeding out "ones-and-dones" and other infrequent visitors—those who visited 5 times or less during our entire observation period. Text preprocessing of the

[4]One of the 25 largest newspaper by circulation in USA. Website: (http://www.bostonglobe.com)

news stories was performed using a standard pipeline from the NLTK toolkit (Bird, 2006). We performed: standard tokenization, stemming, lowercasing, removal of punctuation and determiners.

Our final dataset tracks 9,619,643 visitors over a 174 week period, leading to a total of 90,040,421 non-zero person-week observations. Table 1 shows the summary statistics of our dataset. Of the total visitors, about 97% were from USA. About 7.3% of the unique visitors in the dataset changed their subscription status sometime during the observation period i.e. they either subscribed if they were not subscribed or they canceled if they already had a subscription. Of the remaining, around 87% of the visitors were anonymous, and about 5.7% visitors stayed subscribers throughout.

As is typical for e-commerce businesses, Boston Globe also counts each hit to their website as a unique visit and a typical visit session lasts for 30 minutes i.e. a visitor who spent 45 minutes on the website would have 2 visits attributed to them. Once a visitor clicks on a given news story, that article is counted as read/accessed by them. Globe's users fall into two categories: subscribers and anonymous visitors. Subscribers enjoy unfettered access to news and can be uniquely identified. On the other hand, anonymous visitors are identified via cookies. So it is entirely possible that the same anonymous visitor accessing Globe website using different browsers and computer operating systems would be counted as multiple unique users in our dataset. We understand that this is not an ideal scenario but this is a shortcoming of all cookie-based digital fingerprinting schemes.

	Min.	Median	Mean	Max.
Visits (overall)	1	7	14.9	1150
News articles (overall)	1	9	33.8	47107

Table 1: Summary statistics of the visitation and reading behavior of the visitors to the Globe website.

3 Brief Review: Neural Networks for Customer Base Analysis

Neural Networks (NN) are a powerful class of computational learning models. There are several classes of architectures of NNs including feed-forward neural networks, autoencoders, recurrent neural networks, Generalized Adversarial Networks (GANs) and

others. All these architectures have different modeling strengths and are more suitable for certain prediction tasks compared to others. When neural networks have more than one hidden layer they are called deep neural networks or simply deep learning models. The full discussion of their relative merits is beyond the scope of this paper, so we direct the readers to an updated reference on research in this field (Goodfellow et al., 2016).

Recently, NNs have shown immense promise by achieving state-of-the-art performance on several complex supervised and unsupervised (i.e. lack of labeled data) learning tasks in domains as diverse as speech processing (Hinton et al., 2012), natural language processing (Goldberg, 2016), computer vision/image recognition (Krizhevsky et al., 2012) & reinforcement learning (Mnih et al., 2015). There are several reasons for their superior performance and broad success. First, they are able to learn adaptive basis functions which lead to highly discriminative and efficient non-linear *data representations*. Second, they are able to easily integrate pre-trained embeddings in the form of an *empirical* prior. Finally, they are fast to estimate.

In order to shed further light on the properties of NNs it might be helpful to draw an analogy with probabilistic modeling procedures typically employed in marketing literature. In probabilistic modeling, one assumes a set of random variables some of which can be latent. The researcher can then potentially assign meaning to these random variables based on the context of the modeling task. Finally, inference is performed by "summing over" or "integrating out" the latent variables. The outputs of probabilistic models typically lend themselves to interpretable solutions as mixture distributions over variables of managerial relevance. In contrast with that in neural networks we proceed by assuming a highly flexible class of prediction functions which typically map to low-dimensional spaces. These low dimensional projections might not have overt meaning ascribed to them but their expressivity in distilling the signal from the data gives them superior predictive ability. In spite of their superior predictive ability and other relative merits as discussed above, neural networks face a common criticism due to a lack of interpretability of their outputs and the inherent "black-box" nature of their predictions.

Our approach draws on the best properties of both the neural networks and probabilistic models. In addition to sharing the various attractive properties of neural networks discussed above, our neural network framework also estimates interpretable dynamic customer trajectories. This interpretability is achieved by employing a novel dictionary learning procedure as part of our model which estimates time-varying distributions over a set of latent customer interests. Broadly, the neural architectures that we propose for

modeling consumption dynamics and for estimating customer churn fall under the umbrella of autoencoders and feed-forward neural networks respectively (Goodfellow et al., 2016). So, before we describe our model specifications it will be helpful to give a brief overview of these class of neural networks.

Feed-forward Neural Networks: Feed-forward neural networks (Rosenblatt, 1961; Rumelhart et al., 1985) are one of the oldest class of neural networks. They are potent computational devices and have been shown to be universal function approximators (Cybenko, 1989) as they can approximate families of functions including all continuous functions and functions mapping from one finite dimensional discrete space to another, with any desired non-zero amount of error.

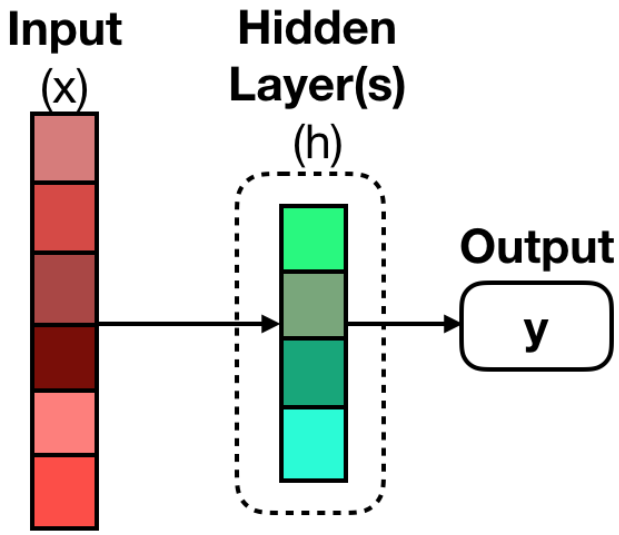

Figure 1: (Best viewed in color) A simple feed-forward neural network with 1 hidden layer. For simplicity, the figure shows only a single arrow between the layers. But in reality the estimable parameters are matrices **W** of size $d_{in} \times d_{out}$, where d_{in} and d_{out} are the dimensionalities of the input layer and the next layer respectively.

It might help to think of feed-forward neural networks as a function NN(x) that takes a d dimensional vector $\mathbf{x}$ as input and produces a m dimensional output vector $\mathbf{y}$. Often this function is complex and almost always it's non-linear e.g. sigmoid $\left(\frac{1}{1+e^{-x}}\right)$, hyperbolic tangent (tanh), ReLu (max(0,$\mathbf{x}$)) etc. Figure 1 shows a typical feed-forward neural network (multilayer perceptron). Neurons are represented by squares and incoming and outgoing arrows represent that layer's inputs and outputs respectively. Each row of neurons can be thought of as representing a vector. For instance, in the given example the input layer is a 6 dimensional vector $\mathbf{x}$, the hidden layer is a 4 dimensional vector,

and finally the output layer is a binary (0/1) scalar $\mathbf{y}$. Each fully connected layer can be thought of as a linear transformation as it implements a vector-matrix multiplication of the form $h = \mathbf{xW}$ (with often a bias term $\mathbf{b}$ added). The weights represent the parameters of the model with each entry $\mathbf{W}_{ij}$ representing the strength of the connection from the i^{th} neuron in the input layer to the j^{th} neuron in the next layer. The values are then transformed by a non-linear function σ before being passed onto the next layer. Following this line of logic, the entire computation performed by the feed-forward neural network in Figure 1 can be written compactly as $\sigma(\mathbf{xW}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$, where $\mathbf{W}_1$ is a 6×4 matrix, $\mathbf{b}_1$ is a 4×1 vector, $\mathbf{W}_2$ is a 4×1 vector, and $\mathbf{b}_2$ is a scalar.

Autoencoders: Autoencoders are similar to feed-forward networks in many respects (cf. Figure 1), except that there is no correct label $\mathbf{y}$ (or supervision). Rather, they just try to predict/reconstruct the input itself. They are allowed to reconstruct the input by using only a limited number of dimensions to represent that reconstruction. The hope is that by allowing them to have only a limited number of dimensions to represent the input data, they will learn to prioritize certain aspects of the input and in the process will learn useful data representations.

Drawing an analogy with traditional statistical methods, feed-forward neural networks can be thought as analogous to ordinary least squares (OLS) regression while autoencoders can be seen as resembling principal component analysis (PCA). Just like PCA autoencoders perform reduced dimension reconstruction of the data, but can employ several layers of non-linearities.

3.1 Neural Network Embeddings

The idea of using neural networks to model high-dimensional discrete distributions such as text data dates back to Bengio et al. (2003), who build a neural probabilistic model for natural language. A key component of any neural network that models such high-dimensional data is to first assign low-dimensional projections, also known as embeddings, to the input data (cf. Figure 2). Embeddings typically constitute the first layer of a neural network and essentially are a mapping from the high-dimensional (one-hot encoded) space of the input data to a low (typically 50-100) dimensional space. Once this mapping from high-dimensional inputs to embeddings has been done, the estimation of the rest of the neural network proceeds in this low-dimensional space. It might

be helpful to think of an embedding layer as a matrix $\mathbf{E} \in \mathbb{R}^{n \times k}$, where n is the (high) dimensionality of the input and k is the (low) dimensionality of the output.

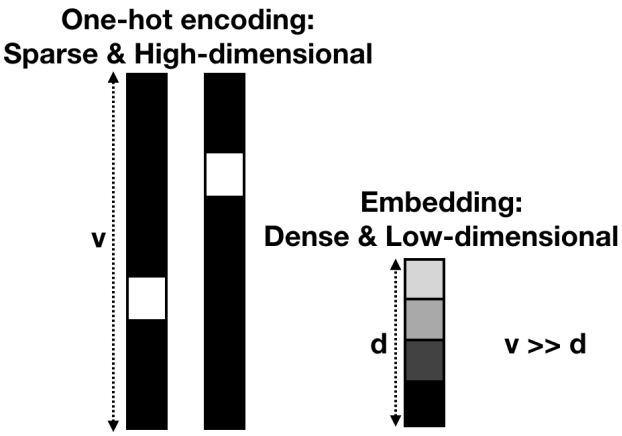

Figure 2: Vector representations and embeddings. The colors represent the potential values of that entry. Black color represents a 0, white represents 1. Other colors represent any real numbered value other than 0 or 1.

The embedding layer also represents one of the comparative advantages of using neural network approaches to model text data as opposed to probabilistic models like Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The low-dimensional embeddings can be estimated completely independently of the estimation of the neural network and used as a data driven (*empirical*) prior. For instance, it is fairly common to use word2vec embeddings (Mikolov et al., 2013b,a) estimated on a far bigger external text corpus as plug-in word embeddings in the embedding layer. That said, if the application demands, one can refrain from using external plug-in embeddings and instead estimate the embeddings along with the rest of the model.

In our setting, the inputs i.e. customers and their content preferences both live in very high dimensional spaces. Their one-hot encoded representation represents them as a vector as big as the total number of customers in our dataset or a vector the size of the vocabulary in English language ($\sim 300,000$), with all 0s and a single 1 at the index corresponding to the lexicographic ordering of these customers/words. Hence, we assign embeddings to both customers and the content in the first layer of our neural network as we will see later.

3.2 Dictionary Learning

As mentioned before, a common criticism of neural network based approaches is the "black-box" style nature of the predictions made by them and the associated lack of interpretability of the results. Since we are interested in modeling customer behavior the managerial interpretability of results is important. To address this concern, we combine neural networks with a popular method for learning parsimonious data representations from signal processing domain known as dictionary learning (Elad and Aharon, 2006; Mairal et al., 2009).

Dictionary learning aims to find a representation (often times sparse) of the input data ($\mathbf{X}_{n\times d}$) in the form of a linear combination of the k items of a dictionary ($\mathbf{D}_{d\times k}$) which are known as *atoms*. These atoms are not required to be orthogonal and they may be over-complete—the dimensions of the representation (k) can be higher than the dimension of the input signal itself (d). These k items/atoms can be seen as analogous to topics in topic models such as LDA and allows us to associate interpretable probabilities to items belonging in certain "topics" similar to the topic-word probability distribution estimated by LDA. The typical objective function optimized by a dictionary learning algorithm is $\text{argmin}_{D,R} \|\mathbf{X} - \mathbf{DR}\|^2$, where $\mathbf{D}$ is the dictionary and $\mathbf{R}$ is the estimated weighting function. The entries of the dictionary are also the parameters of the model and are estimated from data.

Based on this idea, we estimate a customer preference dictionary which models the content consumed each time-period by a customer as a distribution over the atoms of the dictionary. In our case, we choose $k < d$ (k$\sim 30 - 50$) i.e. we represent preferences as a linear combination of an under-complete dictionary. As an illustrative example, suppose the dictionary has three atoms representing *Politics*, *Sports*, and *Business* content, then we use the customer's browsing data to estimate a customer specific distribution over these atoms each time period. For instance, the first time period distribution could be (20%, 40%, 40%) while the second time period distribution could be (55%, 25%, 20%) , and so on.

4 Model Specifications

Building on the background described in the previous sections, we propose a neural network framework for analyzing and profiling the customer base of a firm. First, we

propose an unsupervised neural network (an autoencoder) for modeling the evolution of customers' content consumption behavior over time. Next, we extend this neural network architecture to a supervised learning setting and use it to predict engagement metrics—customer churn and return visits. Before we delve into the details of the models, we describe our model setup.

Model Setup: Let $i = 1, \ldots, n$ index the n customers in our dataset. These customers consume content c_{it} in time-period t, where $t = 1, \ldots, \tau$. The content vector is operationalized as a set of tokens $\{w_1, w_2, \ldots, w_k\}$ over a vocabulary of a total v unique tokens. This could represent any set of traces of a customer's online content consumption activity, for instance, it could be the set of URLs a customer visits online or it could be the text of the articles they read on a website[5]. For the application demonstrated in this paper the content is the headlines of the articles read by the customers at Boston Globe's website. Each customer's time-invariant attributes are denoted by vector a_i and include information such as *zip-code*, *country*, *mobile/desktop*, and *operating system (OS)*. Both the user attributes and the content are represented as one-hot encodings. The subscription status of the customers and the number of visits they made each time period are our dependent variables of interest; for the simplicity of exposition we represent both these customer relationship metrics as y_{it}.

Operationalizing Textual Content: We model the content consumption activities of our panel by modeling the headlines of the articles/stories read by the users. The headlines serve as a good proxy for the actual content body of the article since we are interested in capturing the "topical" nature of the consumers' content consumption.

The most obvious approach to codify the text involves just using the actual words in the text of the headline (also known as unigrams or 1-grams) coded as one-hot vectors over a typical vocabulary of English language ($\approx 300,000$). For instance, in the news story headline "*Johnson talks bitcoin in her first major speech as Fidelity chairman*", the unigram representation would represent all the unique words {*Johnson, talks, bitcoin, in, her, first, major, speech, as, Fidelity, chairman*} as a bag-of-words (i.e. ignore all the ordering information). However, it has a potential shortcoming as it may not entirely capture the semantics (or the meaning) of the news story. To address this problem, at a least to some

[5]In this paper we focus on textual content but our modeling approach is also applicable to other kinds of digital content such as music and video. In such a case, the tokens $\{w_1, w_2, \ldots, w_k\}$ could represent features extracted from audio or video data.

degree, it is common to use extended n-gram representations (bigrams or trigrams) that to some degree preserve the semantic structure of the text. For instance, in the above example, the bigram bag-of-words representation would be {*Johnson-talks, talks-bitcoin, bitcoin-in, in-her, her-first, first-major, major-speech, speech-as, as-Fidelity, Fidelity-chairman*}, similarly the trigram representation will be {*Johnson-talks-bitcoin, talks-bitcoin-in, bitcoin-in-her, in-her-first, her-first-major, first-major-speech, major-speech-as, speech-as-Fidelity, as-Fidelity-chairman,...*}.

Though, the bigram and trigram representations do significantly blow up the vocabulary space as now every two or three word sequence of words is a new word by itself, the increased vocabulary size is not problematic as one gets most benefit from neural networks when the dimensionality of the input data is very high. Hence, in our modeling and analysis we use all 1-grams, 2-grams, 3-grams of the words in the headlines of the articles as the vocabulary to represent the content items c_{it}.

4.1 Modeling Preference Dynamics

To model the dynamics of content consumption, we employ an unsuprvised neural network architecture as shown in Figure 3. We begin by associating d dimensional real-valued embedding vectors v_{c_t} and v_a with the content (time-variant) and the time-invariant user attributes respectively. The content embeddings v_{c_t} capture the evolving nature of a customer's content consumption and are the rows of a matrix $\mathbf{C}$ of dimensions $v \times d$, where v is the number of unique words in the vocabulary. Similarly, v_a captures the time-invariant aspects of a customer's profile and these vectors are the rows of a $n \times d$ matrix $\mathbf{U}$. Since we operationalize content as an unordered set of n-grams, the content embedding v_{c_t} is derived as the average of the embeddings of all the n-grams constituting that content item, as is standard practice. For instance, consider the example shown in Figure 3, where the input content is the set of headlines *"How Danny Ainge is making the Celtics great again"* and *"GE unveils striking new headquarters for Fort Point"*. c_{it} is the vector of all the 1-grams, 2-grams and 3-grams derived from this text; let's call it $\{w_1, w_2, \ldots, w_e\}$. Then the content embedding vector is:

$$v_{c_t} = \frac{1}{e} \sum_{w=1}^{e} v_w,$$

where v_w is the embedding for the n-gram w.

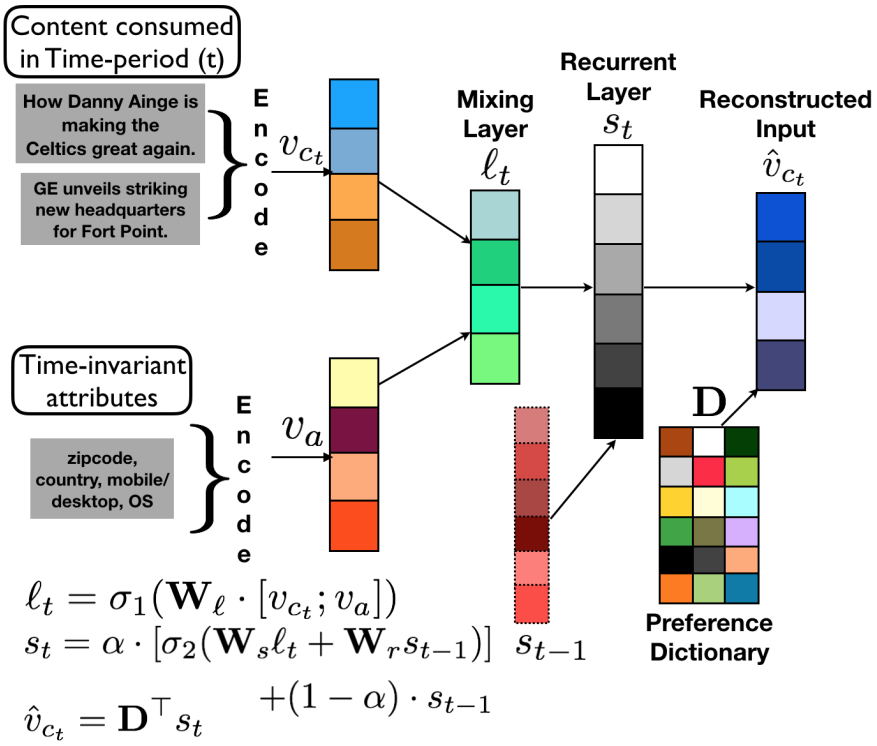

Figure 3: (Best viewed in color) Neural Network Architecture for modeling preference dynamics.

Next, we mix the content and attribute embeddings by concatenating them and feeding them into a standard feed-forward layer with a non-linear activation to obtain an estimate of the hidden state ℓ_t.

$$\ell_t = \sigma_1(\mathbf{W}_\ell \cdot [v_{c_t}; v_a]), \tag{1}$$

where $\sigma_1(x) = \max(0, x)$. We chose ReLu ($\max(0, x)$) non-linear activation owing to its simplicity and its robustness to gradient saturation issues.

The hidden state that we just obtained is an estimate of customers' latent preferences. Since the preferences evolve over time, the next layer in our neural network is a recurrent layer which constructs the updated estimate of the hidden state by combining the previous layer's output from the current time-period (ℓ_t) and current estimate of the state from the previous time-period (s_{t-1}),

$$s_t = \sigma_2(\mathbf{W}_s \ell_t + \mathbf{W}_r s_{t-1}) \tag{2}$$

One potential problem with the above state estimate s_t is that we do not get a smooth

estimate of the state from one time-period to the next, so to alleviate this concern we add exponential smooths with smoothing rate α. Exponential smooths also enable us to have geometrically decaying impact of content consumption i.e. more recently consumed content has a higher contribution towards the estimated hidden state than the content consumed in distant past.

$$s_t = \alpha \cdot [\sigma_2(\mathbf{W}_s \ell_t + \mathbf{W}_r s_{t-1})] + (1 - \alpha) \cdot s_{t-1} \tag{3}$$

We use the recurrent hidden state estimates to get a weighting over the items of the preference dictionary matrix $\mathbf{D}$ by using the softmax $\left[\sigma_2(x) = softmax(x) = \frac{\exp(x)}{\sum_{i=1}^{l} \exp(x_i)}\right]$ as the activation function. This allows us to represent each customer's content preferences as a linear combination of the atoms of the dictionary by converting them to probabilities.

Finally, we reconstruct the input content embedding vector $\hat{v}_{c_t}$ as a weighted linear combination of the atoms as

$$\hat{v}_{c_t} = \mathbf{D}^\top s_t \tag{4}$$

The loss function reconstructs its input as is typical for autoencoders. It can be written as:

$$\begin{aligned} \Theta^* &= \underset{\Theta}{\operatorname{argmin}}\, \mathcal{L}(v_{c_t}, \hat{v}_{c_t}; \Theta) \\ \textit{where} & \\ \Theta &= \{\mathbf{W}_\ell, \mathbf{W}_s, \mathbf{W}_r, \mathbf{D}\} \end{aligned} \tag{5}$$

The dynamic trajectories of evolving customer preferences are just the smoothed hidden state estimates s_t for each time period $t = 1, \ldots, \tau$ for all the n customers. These concise summaries of the customers' content consumption patterns constitute the dynamic customer profiles. Next, we describe the operationalization of the loss function $\mathcal{L}(\cdot)$.

Loss Function: We want to choose a loss function that would make the reconstructed content embedding vector as similar as possible to the input embedding vector. An obvious choice for the loss function is the mean squared error loss, which in this case would become $\mathcal{L}(v_{c_t}, \hat{v}_{c_t}; \Theta) = \|v_{c_t} - \hat{v}_{c_t}\|^2$. However, there is a potential problem with

this loss function as a trivial solution which sets all the model parameters equal to 1 would minimize the loss function by making the output $\hat{v}_{c_t}$ exactly the same as the input $\hat{v}_{c_t}$. A standard way of alleviating this problem and similar edge cases is to artificially add some noise to the estimation procedure. In the context of neural networks one way of doing this is by negative sampling (Mikolov et al., 2013b) which in our case would involve pairing the reconstructed output $\hat{v}_{c_t}$ with m incorrect "negative" inputs $v_{\bar{c}_t}$, where $c_t \neq \bar{c}_t$. Hence, we use a max-margin based ranking loss function, similar to the one used in support vector machines (SVM) (Murphy, 2012) as well as in similar natural language processing (NLP) applications (Socher et al., 2014).

$$\mathcal{L}(v_{c_t}, \hat{v}_{c_t}; \Theta) = \sum_{t=1}^{\tau} \sum_{\bar{c}=1}^{m} \max(0, 1 - \hat{v}_{c_t} \cdot v_{c_t} + \hat{v}_{c_t} \cdot v_{\bar{c}_t}), \tag{6}$$

where $v_{\bar{c}_t}$ represents the m negative samples. In our empirical analysis in this paper we set $m = 10$.

Intuitively, the loss function encourages the correct reconstruction of the input content embedding by having a high inner-product ($\hat{v}_{c_t} v_{c_t}$) and encouraging a low inner-product for the incorrect pairs ($\hat{v}_{c_t} v_{\bar{c}_t}$). Hence, the parameters that are estimated are such that the difference between the correct and incorrect pairings is maximized.

4.2 Modeling Customer Relationship metrics: Churn and Future visits.

Next we turn to the problem of modeling engagement metrics for the customers. We predict the churn probability and the number of future visits that customers will make. The architecture of the neural network is shown in Figure 4 and it is largely similar to the neural network for modeling preference dynamics described in the previous section. The key difference in this model architecture arises from the fact that predicting engagement metrics is a supervised learning problem. Hence, rather than reconstructing the input itself, here we minimize the binary prediction error in the final layer. Additionally, we use the cross-entropy loss function which is commonly used in the literature for these kind of prediction problems, and is defined as:

$$\mathcal{L}(y, \hat{y}; \Theta) = -y \log(\hat{y})$$

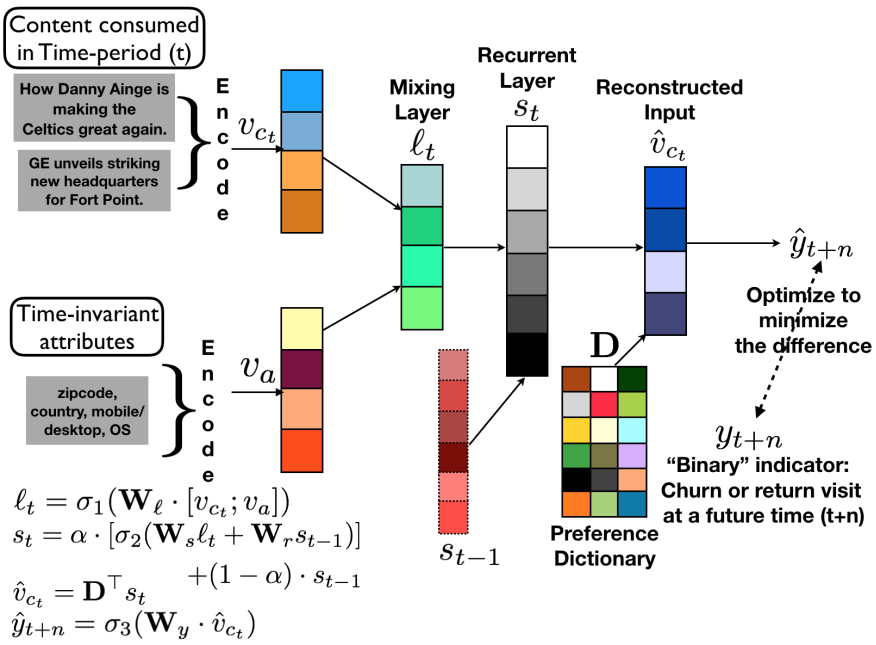

Figure 4: (Best viewed in color) Neural Network Architecture for modeling churn and return visits.

It's easy to see that this is a sensible loss function. When the correct label is predicted $\hat{y} = 1$, the loss becomes 0 and on the other hand when the prediction is bad (say) $\hat{y} \approx 0.05$, the penalty is high. After the reconstruction layer we calculate the estimates of the churn propensity and visits in future time-periods as:

$$\hat{y}_{t+n} = \sigma_3(\mathbf{W}_y \cdot \hat{v}_{c_t}) \tag{7}$$

where $\hat{y}_{t+n}$ are n time-period ahead binary churn/return visit indicators. Our choice of non-linearity (σ_3) also differs from before. Rather than the ReLu function we employ the sigmoid function which converts real numbers into binary class probabilities. Putting it all together, our optimization problem can be written as:

$$\begin{aligned} \Theta^* &= \underset{\Theta}{\operatorname{argmin}}\, \mathcal{L}(y_{t+n}, \hat{y}_{t+n}; \Theta) \\ \textit{where} & \\ \Theta &= \{\mathbf{W}_\ell, \mathbf{W}_s, \mathbf{W}_r, \mathbf{W}_y, \mathbf{D}\} \end{aligned} \tag{8}$$

4.3 Model Estimation & Optimization

The likelihood function of our neural network is non-convex. That is, the optimization procedure may get stuck in a local minimum or a saddle point. Hence starting from different initial points (e.g. different random values for the parameters) one may get different results. Thus one needs to be careful in the optimization of parameters.

At a high level, the estimation procedure of neural networks proceeds in a similar fashion as other statistical methods. Basically, one computes an estimate of the model error over the dataset, then computes the gradient of the parameters with respect to the error, and finally moves the parameters in the direction of the gradient. Different estimation procedures differ in the details of this procedure.

Minibatch stochastic gradient descent (m-SGD) is the gold standard for estimating the parameters of feed-forward neural networks due to its speed and strong convergence properties. It creates small-sized batches (typically 10-20) of the dataset observations and iterates over them repeatedly, each time computing gradients for only those set of observations and updating the relevant model parameters.

The gradients of the parameters for the neural networks are computed efficiently using a procedure known as *backpropagation* (Rumelhart et al., 1985)—which is just another name for methodologically computing the derivatives of a complex expression like our likelihood function in Equation 5 using the chain rule, while caching intermediary results. It involves making a forward pass over the network, computing the errors and then propagating them back to the input layer to update the parameter weights.

We estimated the parameters of our models (Equations 5, 8) using m-SGD with backpropagation using the TensorFlow library (Abadi et al., 2016). η, the learning rate of m-SGD, was chosen as the value that minimized the loss function on a separate held-out development dataset (1 month of user activity).

We estimated our models with the input dimensionality of embeddings $d = 100$ and pre-trained GloVe embeddings[6] for v_{c_t}. The attribute embeddings v_a were initialized uniformly at random. All other parameters of the model were also initialized uniformly at random as is common practice. The model estimation was done for 300 epochs till the convergence criteria was met.

[6]Available for download from: http://nlp.stanford.edu/data/glove.6B.zip

5 Results

In this section we show the empirical performance of our models in capturing the dynamics of customers' interests and on predicting customer relationship metrics. Our results are divided into two parts. First, we show the trajectories of customer content interests qualitatively at both the population and individual-level. Next, we turn to quantitative evaluations. There, we first illustrate the economically significant improvement in churn prediction over competing methods. Finally we turn to predicting customer engagement metrics using our dynamic neural network architecture.

5.1 Qualitative Results

We estimated the neural network we described in the previous section (Equation 5) to estimate evolving customer interests. The rows of the preference dictionary **D** describe customer interests akin to topics in topic-modeling approaches like LDA. Though, unlike topic-models, the dictionary encodes interests as real-valued numbers as opposed to a set of discrete word clusters. However, one can easily find the words associated with a given row of the dictionary matrix by simply computing the nearest neighbor vectors of that dictionary item with the input word embeddings v_{c_t}.

In our estimation we chose the dimensionality of the preference dictionary as k=30. Sample atoms/descriptors from the estimated preference dictionary are given below:

- *obamacare, clinton, trump, trickle-down, voter-ballot*
  (Topic: Politics/Elections)
- *olympics, deflategate, ortiz, judge, brady-peyton, patriots*
  (Topic: Sports-1)
- *beach, summer, snow, sharks, chatham, clams, surf*
  (Topic: Travel/Vacation-1)
- *accelerators, inflation, infrastructure, driverless, blockchain, bond-rating*
  (Topic: Finance/Business/Technology)
- *mansion, fritter, chowder, montauk, trail, beachhouse, ferry, shower*
  (Topic: Travel/Vacation-2)
- *eagles, sandoval, yankees, pitch, bullpen, defeat, nets, caroline*
  (Topic: Sports-2)

As can be seen, the dictionary does seem to correspond to intuitive categories of customer preferences. The topical content shown above broadly captures information regarding *Politics/Elections*, *Sports*, *Travel/Vacation*, *Finance/Business/Technology*, *Travel/Vacation*, and *Sports*, respectively. It is perfectly normal for a topic to be a mixture of two closely related concepts e.g. travel, vacation, and conversely, to have more than one topic be assigned to a similar concept e.g. sports in the case above. Note that the topic names mentioned above e.g. politics, elections, sports are assigned manually by inspection by looking at the composition of the words that comprise that topic. Our model just outputs a clustered collection of words.

The individual-level customer trajectories given by the vector s_t provide weightings over these topics at a given time-point. These vectors can be averaged over the entire population to get population-level trajectories of different topics. These highlight broad population-level trends in content consumption. Figure 5 shows such population-level visualizations of the trajectories. As can be seen, they reveal several salient trends in consumption e.g. a general upward trend in politics content consumption around the time of the 2016 US Presidential election; a population-wide increase in sports consumption around the time that New England Patriots made the NFL playoffs[7].

Figure 6 shows the evolving interests of four randomly chosen individuals in our dataset who made recurring visits. These plots show the s_t vector i.e. the evolving weighting over the preference dictionary for those individuals. As can be seen, for some individuals the interests stay constant over the entire observation period e.g visitor 2 (top-right) and visitor 3 (bottom-left), for other changes they change gradually e.g. visitor 4's (bottom-right) interests in politics and sports change gradually around the time of the start of 2016 US Presidential election primaries. Some visitors e.g. visitors 3,4 (bottom) show strong periodical variations in their interest in content on lifestyle/travel.

These visualizations provide a proof-of-concept that our model is capturing nuanced aspects of the temporal evolution of the customer interests and preferences. These time-series plots can serve as a dashboard for marketing analysts to track the ebb-and-flow of customer preferences. In the next section we take our analysis a step further and use our neural network to predict customers' relationship metrics—churn and return visits.

[7]It is worth pointing out that a significant fraction of the visitors to Boston Globe are from New England area and support Boston sports team, so that is why we see this trend. This might not be true in general of sports news consumption.

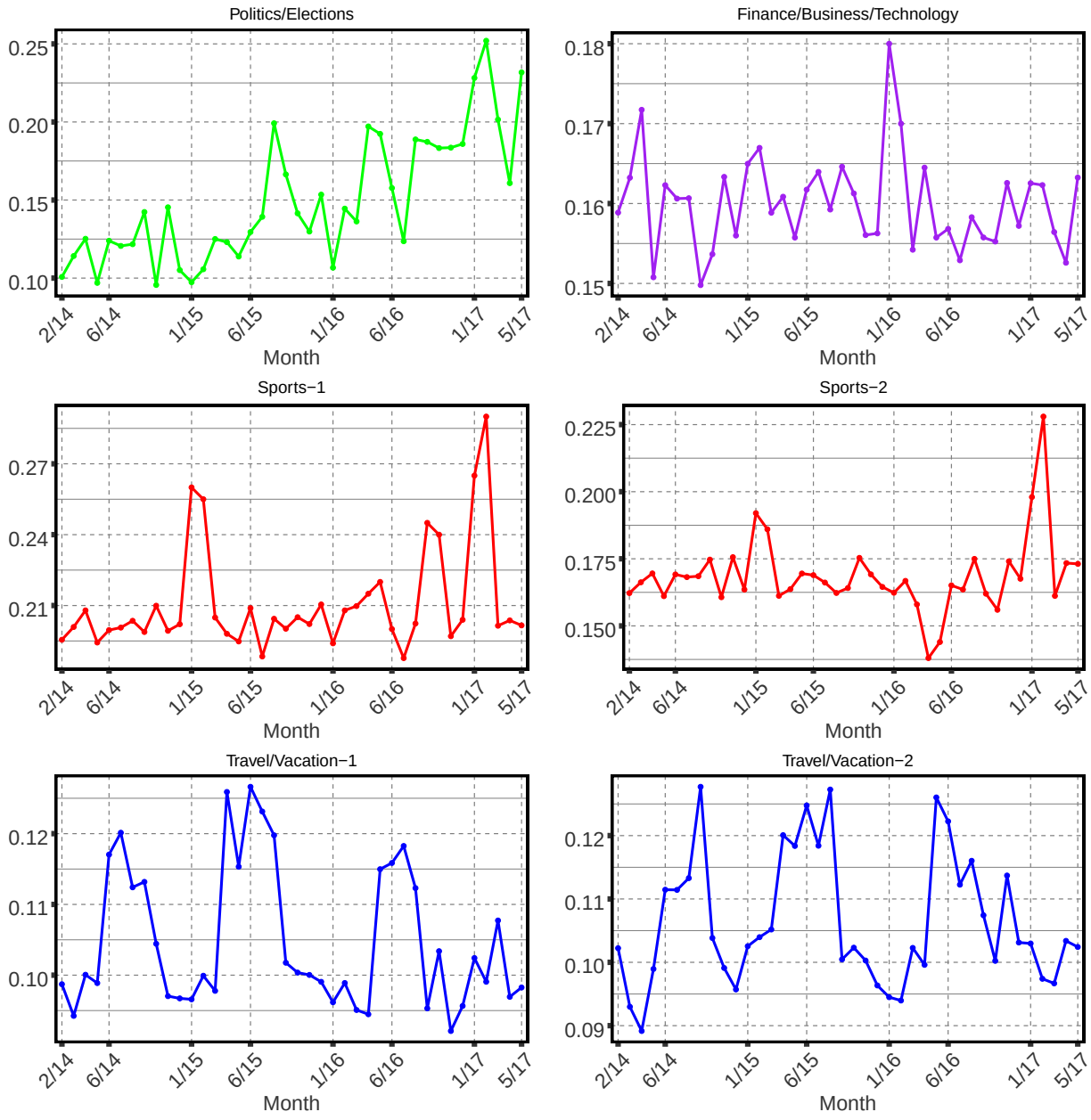

Figure 5: Figure showing population-level trends for Politics/Elections, Sports, Finance/Business, Travel/Vacation content consumption by Globe readers over a 3-year period. *Note:* The y-axis represents the corresponding weighting learned by our model via the s_t vector. The plots shown above correspond to the topical content of these topics shown earlier in this section.

Figure 6: (Best viewed in color) Figure showing individual-level content consumption trends over a 3-year period for 4 randomly chosen Globe readers. *Note:* The y-axis represents the corresponding individual weighting learned by our model via the s_t vector. The plots shown above correspond to the topical content of these topics shown earlier in this section.

5.2 Quantitative Results

Next we use our neural network described in Equation 8 to predict whether a customer will churn in the hold-out period. Further, we also predict the number of visits made by the customers' in the hold-out period.

Baseline Methods: The baseline methods that we compare our approach against comprise a range of alternatives.

1. *Demographics*: Using the zip-code, state, country, user-id as covariates in a predictive classifier with the exact same loss function (cross-entropy) that our neural network used.

2. *Raw Traces*: Using the high dimensional content as covariates in a predictive classifier with the exact same loss function (cross-entropy) that our neural network used.

3. *Pooled Neural Network*: Here we estimate our model but with no time dimension. We pool all the content consumed by a user over the entire observation period i.e. no time dimension and then input it to our neural network.

4. *Latent Dirichlet Allocation (LDA):* LDA (Blei et al., 2003) is a popular probabilistic Bayesian model used for text analysis. It has been used to model user-generated content (UGC) in several marketing applications (Büschken and Allenby, 2016; Liu and Toubia, 2017). If we are given a set of documents, each containing set of words, then LDA models each document as a mixture of small number of latent topics and further attributes each word's creation to one of these topics. These estimated topics can be thought of as clusters of similar words. LDA was not proposed for a setup like ours where we observe the longitudinal behavior of a set of users. So, we adapt LDA to our setting by pooling all the content consumed by a given user i.e. removing the temporal component and further considering each user as a document. Finally, we use the LDA topics as covariates in a predictive classifier with the exact same loss function (cross-entropy) that our neural network used.

Notice that baseline methods are ordered in terms of increasing complexity. Here, we want to understand how exactly different elements of our model contribute towards predictive accuracy. So, first we establish that preferences contain more predictive signal than just user demographics. Next, we show that a neural network learning the low-dimensional projections adds predictive power compared to just using sparse high-dimensional covariates. As a natural next step, we show that it is essential to model the dynamics of content consumption and that using only cross-sectional variation leads to loss of predictive accuracy. Finally, we highlight the improved performance obtained by our dynamic neural network over a state-of-the-art text modeling method.

5.2.1 Churn Prediction

One important question for any subscription-driven online content provider is to predict which subscribers are about to churn. The ability to correctly identify such individuals can allow platforms to do pro-active churn management by targeting discounted offerings towards such individuals (Ascarza et al., 2018). Towards that end, we used our model to make forward-looking churn predictions of Boston Globe's online user base. In

particular, we make "binary" churn predictions in the next 1 and 5 months. Alternatively, we could have formulated the prediction problem as one of survival analysis[8]. Though, both of these approaches have widespread use, we chose the former over the latter owing to simplicity of exposition, especially since the focus of this paper is modeling dynamic consumption patterns rather than optimizing pro-active churn management.

We report our results in Table 3 and the confusion matrices for the predictions are reported in Tables 4 and 5. Specifically, we report the precision, recall, and F1-scores (Provost and Fawcett, 2013) of various methods. These metrics are machine learning concepts that provide a more nuanced look into the performance of predictive models especially when there is a skew in the distribution of class labels. For a binary classification problem, as is our case, there are four possible scenarios illustrated in Table 2

True Label	Predicted Label	Type
1	1	True Positive
0	0	True Negative
0	1	False Positive
1	0	False Negative

Table 2: Predictions of a binary classifier.

Given this, precision, recall, and F1-score are formally defined as:

$$\text{Precision} = \frac{\text{Number of True Positives}}{\text{Number of True Positives + Number of False Positives}}$$

$$\text{Recall} = \frac{\text{Number of True Positives}}{\text{Number of True Positives + Number of False Negatives}}$$

As we can see, precision is related to Type I error rate while recall is related to Type II error rate. Looking at the results, our method performs at least 8 percentage points better in both precision and recall compared to the next best approach. In our context, improving precision can lead to more effective pro-active churn management by reducing the number of incorrectly targeted discounts. On the other hand, improving recall leads to correctly targeting discounts to individuals who actually churn. Hence, lesser lost revenue due to canceled subscriptions. It is a little challenging to translate the improved performance into dollar-terms for a couple of reasons. First, the revenue impact depends on the effectiveness of the pro-active churn management campaign as some fraction of the customers will churn in spite of the discounts offered to them. Second, even for the customers that we would prevent from churning, we do not have a good

[8]We could have used Cox proportional hazards framework to estimate the "risk" of churn.

estimate of their lifetime dollar-value.

Method	1 month			5 months		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Demographics	0.19	0.21	0.20	0.22	0.26	0.24
Raw Traces	0.41	0.52	0.46	0.49	0.58	0.53
Pooled Neural Network	0.53	0.64	0.58	0.60	0.69	0.64
LDA	0.57	0.66	0.61	0.62	0.71	0.66
Our approach	0.65	0.74	**0.69**	0.70	0.80	**0.75**

Table 3: Predicting user churn in the next 1 month and next 5 months using various methods. Note: Precision= $\frac{\text{\# True Positives}}{\text{\# TruePositives+\# False Positives}}$, Recall= $\frac{\text{\# True Positives}}{\text{\# TruePositives+\# False Negatives}}$, F1-score $= \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision + Recall}}$= Harmonic mean of Precision and Recall.

		Predicted Outcome		
		0	**1**	**Total**
True Outcome	**0**	9,041,177	164,822	9,205,999
	1	107,548	306,096	413,644
	Total	9,148,725	470,918	

		Predicted Outcome		
		0	**1**	**Total**
	0	9,000,048	205,951	9,205,999
	1	140,639	273,005	413,644
	Total	9,140,687	478,956	

Table 4: Confusion matrices showing predictions of our approach (left) and LDA (right) for 1 month look ahead predictions shown in Table 3. Note that "1" represents a canceled subscription (or churn) and "0" represents a subscriber.

However, we can perform a back-of-the-envelope calculation to bound the revenue impact. About 4.3% of the total subscribers in our dataset canceled their subscriptions which amounts to roughly 413,644 customers. As we can glean from Table 4, an improved precision of 8% for one month look-ahead prediction implies that our approach has around 41129 fewer false positives than the next best approach (LDA). Assuming the cost of discount for pro-actively preventing churn as $20 per customer, implies a saving of around $822,580[9] due to more precise targeting. Similarly, an improved recall of 8% implies 33091 fewer false negatives. From our conversations with Boston Globe, we realize the effectiveness of pro-active churn management to be around 30-40% and the average life-time dollar value of a customer to be around $150. So, the net-positive benefit of preventing churn is in the range $1,489,095-$1,985,460[10] and the net cost of

[9] 41129×20

[10] $33091 \times 150.0 \times .30$ and $33091 \times 150.0 \times .40$

True Outcome	Predicted Outcome 0	Predicted Outcome 1	Total
0	9,064,178	141,821	9,205,999
1	82729	330,915	413,644
Total	9,146,907	472,736	

	Predicted Outcome 0	Predicted Outcome 1	Total
0	9,025,997	180,002	9,205,999
1	119,957	293,687	413,644
Total	9,145,954	473,689	

Table 5: Confusion matrices showing predictions of our approach (left) and LDA (right) for 5 months look ahead predictions shown in Table 3. Note that "1" represents a canceled subscription (or churn) and "0" represents a subscriber.

offering discounts is around $661,820[11]. Our approach, therefore, had a net-positive revenue impact in the range $1.65-$2.15 million[12]. A similar calculation for the five month look-ahead prediction suggests a positive revenue impact in the range $1.69-2.25 million. Hence, our results are not only statistically significant but they are also economically significant.

5.2.2 Predicting Visits

In addition to churn management, digital content providers are also interested in maintaining sustained customer engagement and increasing repeat visitors. Increased engagement translates to increased advertisement revenue and also increased propensity for individuals to become subscribers or continue being subscribers. So, we use our model for two prediction tasks related to forward-looking engagement prediction. First, we make a binary prediction regarding whether a visitor will make repeat visits in the hold-out period of next 1 and 5 months. Next, we predict exactly how many visits will that visitor make in the hold-out period using a Poisson loss function. Note that the second prediction problem is much harder, but it is a much finer barometer of customer engagement. The results are shown in Tables 6 and 7 and as can be seen, our neural network model performs significantly better than the competing baseline approaches.

Firms can use our results to improve the accuracies of their individual-level interventions designed to ensure sustained customer engagement. For instance, they can ensure accurate targeting of reminder e-mails to customers with low predicted future engagement

[11] 33091×20

[12] 822,580-661,820+1,489,095 and 822,580-661,820+1,985,460

Method	1 month	5 months
	Classification Error	Classification Error
Demographics	0.46	0.38
Raw Traces	0.33	0.29
Pooled Neural Network	0.28	0.22
LDA	0.25	0.21
Our approach	**0.18**	**0.14**

Table 6: Predicting (binary) visit or not in the next 1 month and next 5 months time horizon. Note: Classification Error = $\sum_{i=1}^{n} \frac{\mathbb{1}_{y_i \neq \hat{y}_i}}{n}$, where y is the correct visit counts, $\hat{y}$ is the predicted visit counts and $n = 9,619,643$ is the total number of observations in our data.

Method	1 month	5 months
	MAE	MAE
Demographics	9.46	7.31
Raw Traces	1.86	1.30
Pooled Neural Network	1.63	1.21
LDA	1.41	1.15
Our approach	**1.06**	**0.93**

Table 7: Predicting number of visits in the next 1 month and next 5 months time horizon. Note: Mean Absolute Error (MAE) = $\sum_{i=1}^{n} \frac{\|y_i - \hat{y}_i\|}{n}$, where y is the correct visit counts, $\hat{y}$ is the predicted visit counts and $n = 9,619,643$ is the total number of observations in our data.

levels.

6 Discussion

This paper developed a dynamic neural network with dictionary learning that can be used for analyzing a firm's customer base and for creating digital profiles of the customers. We applied our modeling approach to study the evolution of consumers' content consumption behavior and to predict their future engagement metrics. Our work contributes substantively by tapping into the all-important problem of opening a window into consumers' dynamic interests. This is not only important for digital marketers but also of immense societal and public policy importance.

Our results highlight the superior ability of our model in capturing nuances in the dynamic consumption patterns. The content trajectories estimated by our model provide concise summaries of customers' interests which can be used to target advertisements or

to recommend articles to read (or more broadly, products). In addition to summarizing customers' content preferences, our neural network's learned low-dimensional projections have sufficient predictive power. Our approach outperforms a host of competitive baseline methods in predicting customers' future engagement metrics. The predictive performance boost achieved by using our approach is not just statistically significant but also economically significant.

Methodologically our novel neural network architecture represents significant advances over extant approaches. Our approach is computationally efficient to estimate and it scales to large data sizes. The preference dictionary learning component of our model ensures that we learn interpretable representations of customers' preferences as a mixture distribution over their latent interests. So, it obviates the need of a Bayesian model for interpretability[13] and hence also sidesteps the need for slow MCMC[14] based inference. The ability to incorporate flexible non-linearities into our model allows us to learn adaptive basis functions that capture highly expressive data representations. Additionally, we also incorporate empirical priors into our model in the form of pre-trained embeddings estimated on external data sources. This is a big comparative advantage of our model as it allows our model to harness external data sources to improve the predictive task at hand. In summary, our modeling framework combined the efficiency and the representational power of neural networks with the interpretability of probabilistic models. To the best of our knowledge, this is the first paper to use fine-granular data to model the dynamics of online content consumption—a predominant way of content consumption these days. This is also the first paper to propose a novel neural network architecture for an empirical problem of relevance to digital marketers and economists.

6.1 Managerial Implications

Our model provides an end-to-end customer analytics framework which can be used by marketing managers to track the health of their customer base as well as to design suitable interventions for retaining customers. To that end, our results have several important managerial implications.

[13]It is fairly common in quantitative marketing literature to build Bayesian models for interpreting the posterior distribution of certain parameters of managerial interest.

[14]Typically, MCMC is used for inference in Bayesian models to sample from an intractable posterior distribution.

Zooming-in on customers: The trajectories of customers' consumption provide concise summaries of the ebb-and-flow of their interests and preferences. The trajectories are easy to interpret as they provide time-varying weights over a set of intuitive mixtures of topical content and they could be part of a marketing analyst's customer analytics dashboard. Marketing managers could compare the population-level consumption trajectories with individual-level trajectories for the same "topic" over time and calculate the divergence of individual-level trajectories over and above than can be explained by population-level trends. They could further zoom-in on individuals with extreme dynamics and track them to see if the patterns sustain. This provides valuable information about the evolution of an individual's content preferences that can be used to target advertisements or recommend products.

Pro-active churn management: The superior predictive capability of our model as shown by the results could lead to improved customer retention. When employed as part of a potential pro-active churn management campaign, our approach can identify potential churners with higher accuracy. As a next step, the firm could offer discounted offerings to those customers, hence reducing the churn. We expect that with the superior ability of our model to identify potential churners, we could narrow the pool of would-be churners who call to cancel their subscription. Hence, having a significant impact on revenue as was also suggested by the back-of-the-envelope calculation in the results section.

Personal Pricing: The ability to correctly estimate each customer's willingness-to-pay (WTP) to offer them targeted prices a.k.a *first-degree price discrimination* is one of the holy-grails of economics (Dubé and Misra, 2017; Rossi et al., 1996; Tirole, 1988). Our approach takes a step towards realizing such personal pricing solutions. The ability of our neural network in capturing nuanced ebbs-and-flows of customer preferences and other individual-level idiosyncrasies suggests that it also captures information that might be highly predictive of their willingness-to-pay (WTP). It can therefore be used to predict customers' WTP and for potentially designing personalized digital paywall solutions as is currently being done by The Wall Street Journal[15].

[15] http://www.niemanlab.org/2018/02/after-years-of-testing-the-wall-street-journal-has-built-a-paywall-that-bends-to-the-individual-reader/

6.2 Customer Preferences & Supply of News

Our modeling framework uses the readership patterns to estimate dynamic customer profiles and to predict future engagement patterns. We model the news consumption behavior from a demand-side perspective by assuming that readership in current time period is only driven by readership in previous time periods. In other marketing contexts such as retailing or consumer packaged goods, this might be a fine assumption. However, in the case of online news consumption, a strong supply-side mechanism exists due to the variation in availability of content. For example, the readers probably consumed more political content during the 2016 US Presidential Election not just because of their content preferences but also due to the increased supply of political news stories during that time period. Hence, the dynamic consumption trajectories estimated by our model might not solely represent customers' preferences but rather represent an interaction of their preferences with the available content.

Disentangling this simultaneity in news consumption and modeling the impact of content availability on consumption patterns is important for furthering our understanding of shifts in customer preferences. Getting reliable estimates of the true underlying customer preferences is also important from a substantive standpoint as it has implications for policy design and for gauging the health of a democracy. It is, though, sufficient to just model the consumption behavior for the predictive modeling applications that we are interested in. The focus of this paper is to provide an efficient framework for marketing managers to perform customer base analysis by allowing them to build accurate predictive models.

6.3 External Validity

A natural question to ask is regarding the generalizability of our findings to other contexts. Our proposed approach is general enough to model any other kind of online data source e.g. various types of user-generated content (UGC), customer purchase history from an online retailer, audio or video content. None of the modeling assumptions that we made or our broad framework had anything specifically tuned to Boston Globe or news consumption more generally. The main difference in modeling other types of data will be regarding the type of features extracted from the input data. For instance, in this paper we extracted word tokens based features from the content consumed by the customers. So, for video or audio data we will have to extract a different set of features that

are relevant for those domains. That said, some of the findings of our model might be specific to Boston Globe's visitor base, for instance, the intense interest in sports news, and might not be be representative of news consumption in general.

6.4 Conclusion & Limitations

We are living in the age of an information deluge. Firms are overwhelming customers with highly intrusive advertisements, emails & coupons since they lack reliable estimates of customer preferences. It is partly due to the companies not being able to efficiently harvest economically significant signal from the large swathes of clickstream data and partly due to their inability in collecting relevant data in the first place (Mela and Moorman, 2018). Customer analytics approaches like ours that are both computationally and statistically efficient have the potential to move the firms towards their goal of tapping into customers' minds and increasing the relevance of their messages (in the form of advertisements, emails & coupons).

That said, our framework is not without limitations. First, we model only one kind of digital footprints left by consumers—content consumption. Future work should model other kinds of data e.g. online search history, comments and other types of UGC. Second, future work could go beyond bag-of-words assumption we made while modeling our content. It could for instance use convolutional neural networks (CNN) or attention mechanisms to model the relative importance of different words in the consumed content. The magnitude of economic impact of these methodological choices is an empirical question and is tough to predict beforehand. Third, we only model the demand-side and assume that consumers consumption patterns are driven only by their consumption in previous time periods. It is an interesting avenue of future research to model the interaction of content availability with readers' consumption preferences. We hope our work will inspire future research to overcome these limitations in pushing the limits of our understanding of the dynamics of online content consumption.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow:

Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

Makoto Abe. "Counting your Customers" one by one: A Hierarchical Bayes extension to the Pareto/NBD model. *Marketing Science*, 28(3):541–553, 2009.

Asim Ansari and Carl F Mela. E-customization. *Journal of marketing research*, 40(2):131–145, 2003.

Eva Ascarza, Scott A Neslin, Oded Netzer, Zachery Anderson, Peter S Fader, Sunil Gupta, Bruce GS Hardie, Aurélie Lemmens, Barak Libai, David Neal, et al. In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Customer Needs and Solutions*, 5(1-2):65–81, 2018.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Steven Bird. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Anand V Bodapati. Recommendation systems with purchase data. *Journal of marketing research*, 45(1):77–93, 2008.

Joachim Büschken and Greg M Allenby. Sentence-based text analysis for customer reviews. *Marketing Science*, 35(6):953–975, 2016.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989.

Ryan Dew, Yang Li, and Asim Ansari. Dynamic preference heterogeneity. 2017.

Jean-Pierre Dubé and Sanjog Misra. Scalable price targeting. Technical report, National Bureau of Economic Research, 2017.

Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

Peter S Fader, Bruce GS Hardie, and Jen Shang. Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science*, 29(6):1086–1108, 2010.

Vivek F Farias and Andrew A Li. Learning preferences with side information. 2016.

Yoav Goldberg. A primer on neural network models for natural language processing. *J. Artif. Intell. Res.(JAIR)*, 57:345–420, 2016.

Avi Goldfarb and Catherine Tucker. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404, 2011.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.

Peter M Guadagni and John DC Little. A logit model of brand choice calibrated on scanner data. *Marketing science*, 2(3):203–238, 1983.

John R Hauser, Glen L Urban, Guilherme Liberali, and Michael Braun. Website morphing. *Marketing Science*, 28(2):202–223, 2009.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Jin Gyo Kim, Ulrich Menzefricke, and Fred M Feinberg. Modeling parametric evolution in a random utility framework. *Journal of Business & Economic Statistics*, 23(3):282–294, 2005.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Mohamed Lachaab, Asim Ansari, Kamel Jedidi, and Abdelwahed Trabelsi. Modeling preference evolution in discrete choice models: A bayesian state-space approach. *Quantitative Marketing and Economics*, 4(1):57–81, 2006.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196, 2014.

Jia Liu and Olivier Toubia. How do consumers form online search queries? the importance of activation probabilities between queries and results. 2017.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 689–696, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553463. URL http://doi.acm.org/10.1145/1553374.1553463.

David Martens, Foster Provost, Jessica Clark, and Enric Junqué de Fortuny. Mining massive fine-grained behavior data to improve predictive analytics. *MIS quarterly*, 40 (4), 2016.

Carl F Mela and Christine Moorman. Why marketing analytics hasn't lived up to its promise. *Harvard Business Review*, 2018.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013b.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Wendy W Moe. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of consumer psychology*, 13(1-2):29–39, 2003.

Wendy W Moe and Peter S Fader. Dynamic conversion behavior at e-commerce sites. *Management Science*, 50(3):326–335, 2004.

Kevin P Murphy. Machine learning: A probabilistic perspective., 2012.

Ramya Neelamegham and Pradeep K Chintagunta. Modeling and forecasting the sales of technology products. *Quantitative Marketing and Economics*, 2(3):195–232, 2004.

Oded Netzer, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko. Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31(3): 521–543, 2012.

Claudia Perlich, Brian Dalessandro, Troy Raeder, Ori Stitelman, and Foster Provost. Ma-

chine learning for targeted display advertising: Transfer learning in action. *Machine learning*, 95(1):103–127, 2014.

Foster Provost and Tom Fawcett. Data science and its relationship to big data and data-driven decision making. *Big data*, 1(1):51–59, 2013.

Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, CORNELL AERONAUTICAL LAB INC BUFFALO NY, 1961.

Peter E Rossi, Robert E McCulloch, and Greg M Allenby. The value of purchase history data in target marketing. *Marketing Science*, 15(4):321–340, 1996.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

David C Schmittlein, Donald G Morrison, and Richard Colombo. Counting your customers: Who-are they and what will they do next? *Management science*, 33(1):1–24, 1987.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, et al. Characterizing geographic variation in well-being using tweets. In *ICWSM*, 2013.

Lauren Sinnenberg, Alison M Buttenheim, Kevin Padrez, Christina Mancheno, Lyle Ungar, and Raina M Merchant. Twitter as a tool for health research: a systematic review. *American Journal of Public Health (ajph)*, 2017.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2(1):207–218, 2014.

Srinivasaraghavan Sriram and Manohar U Kalwani. Optimal advertising and promotion budgets in dynamic markets with brand equity as a mediating variable. *Management Science*, 53(1):46–60, 2007.

Artem Timoshenko and John R Hauser. Identifying customer needs from user-generated content. 2017.

Jean Tirole. *The theory of industrial organization*. MIT press, 1988.

Seshadri Tirunillai and Gerard J Tellis. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, 51(4):463–479, 2014.

Michael Trusov, Liye Ma, and Zainab Jamal. Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting. *Marketing Science*, 35(3):405–426, 2016.

Glen L Urban, Guilherme Liberali, Erin MacDonald, Robert Bordley, and John R Hauser. Morphing banner advertising. *Marketing Science*, 33(1):27–46, 2013.

Hema Yoganarasimhan. Search personalization using machine learning. *Social Science Research Network (SSRN)*, 2016.