Marketing Science

I Will Survive: Predicting Business Failures from Customer Ratings

Christof Naumzik, Stefan Feuerriegel, Markus Weinmann,

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.)
and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual
professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to
transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

I Will Survive: Predicting Business Failures from Customer Ratings

Christof Naumzik,[a] Stefan Feuerriegel,[a,b,*] Markus Weinmann[c,d]

[a] ETH Zurich, 8092 Zurich, Switzerland; [b] LMU Munich, 80539 Munich, Germany; [c] University of Cologne, 50923 Cologne, Germany;
[d] Erasmus University, 3062 PA Rotterdam, Netherlands
*Corresponding author
Contact: cnaumzik@ethz.ch (CN); feuerriegel@lmu.de (SF); weinmann@wiso.uni-koeln.de, https://orcid.org/0000-0002-8342-2756 (MW)

Abstract. The success, if not survival, of service businesses depends on their ability to satisfy their customers. Yet, businesses often recognize slumping customer satisfaction too late and ultimately fail. To prevent this, marketers require early warning tools. In this paper, we build upon online ratings as a direct measure of customer satisfaction and, based on this, predict business failures. Specifically, we develop a variable-duration hidden Markov model; it models the rating sequence of a service business in order to predict the likelihood of failure. Using 64,887 ratings from 921 restaurants, we find that our model detects business failures with a balanced accuracy of 78.02%, and this prediction is even possible several months in advance. In comparison, simple metrics from practice have limited ability in predicting business failures; for instance, the mean rating yields a balanced accuracy of only around 50%. Furthermore, our model recovers a latent state ("at risk") with an elevated failure rate. Avoiding the at-risk state is associated with a reduction in the failure rate of more than 41.41%. Our research thus entails direct managerial implications: we assist marketers in monitoring customer satisfaction and, for this purpose, offer a data-driven tool that provides early warnings of impending business failures.

Keywords: hidden Markov model • customer ratings • business failure • service management

1. Introduction

The success (and survival) of service businesses is highly dependent on the perceived quality of their services (Parasuraman et al. 1988). If quality falls behind expectations, financial figures could decline, which might ultimately lead to failure and business closure. Given that more than 50% of service businesses fail within 4.5 years (Luo and Stark 2015), predicting those failures well in advance would be of enormous practical relevance. If businesses are at risk for failing, marketing managers will require timely feedback—preferably from metrics that directly measure customer satisfaction—so as to be able to adjust their services and operations, ultimately preventing a business failure.

Marketers have several tools at their disposal to gather feedback on business performance (Ittner and Larcker 2003). One example from practice is the monitoring of financial indicators (Sharma and Mahajan 1980, Mahajan et al. 2002), which have inherent deficits because they are only a delayed reflection of poor customer satisfaction (e.g., earnings reports at the end of a quarter; cf. Ittner and Larcker 2003, Lervik-Olsen et al. 2014). Hence, literature recommends to gather direct feedback on customer satisfaction (Ittner and Larcker 2003). A well-known method is SERVQUAL, a survey-based instrument (Parasuraman et al. 1988) that has proven to be very reliable over decades of research (for an overview, see Asubonteng et al. 1996, Buttle 1996, Ladhari 2009). SERVQUAL, with its 22 items, allows for precise assessments of service quality, yet, because of the survey-based method of data collection, prohibits scalable use. To this end, online ratings, which are said to reflect customer satisfaction (e.g., Ho et al. 2017, Schneider et al. 2021), allow for a scalable use

and provide marketing managers with direct feedback regarding online customer satisfaction.

In this study, we use online customer ratings to predict the likelihood of business failures. By doing so, we provide marketing managers an early warning indicator that might help to prevent business failures. Predicting business failures on the basis of customer ratings, however, is a challenge. One might think that studying the most recent ratings or monitoring the average rating of a service inherently reflects service performance and will thus be sufficient. Yet this is untrue, as such metrics lack predictive power. As we show later, the mean rating score yields a balanced accuracy of around 50%, and is thus on par with a random guess. As we will demonstrate, making inferences from customer ratings requires a more sophisticated modeling framework that considers the full rating sequence.

To predict business failures from a rating sequence, we draw upon the hidden Markov model (HMM)-based framework (e.g., Netzer et al. 2008, Schweidel et al. 2011, Ascarza and Hardie 2013, Schwartz et al. 2014, Ascarza et al. 2018). Formally, we model the sequence of individual ratings so that each rating is a noisy realization of a latent state. The latent state refers to the current level of service performance. As we shall see later, we find that one of the latent states—the "at-risk" state—serves as an early warning indicator: it signals phases during which businesses are at an elevated risk of failure. In our work, we model the failure risk as a function of the duration spent in the at-risk state, for which we develop a variable-duration hidden Markov model (VD-HMM). Different from a traditional HMM, our modeling approach allows us to make inferences not only from a latent state but from the duration spent in a latent state.

The intuition of our model is as follows. Let us assume a restaurant, for which we aim to predict whether it will survive or fail in the near future based on its rating history. However, the course of ratings is likely to change over time, as restaurants vary in their service performance. The service performance can further change due to various events such as, for example, the hiring of a new chef, the refurnishing of the restaurant, or the introduction of a new menu. A restaurant that is well rated over a long duration will have many repeat customers and is thus more likely to survive. Conversely, some restaurants' service performance might also not be sufficient to survive, and, hence, they will find themselves in an at-risk state of business failure. Being in such an at-risk state just once is unlikely to force the restaurant to close immediately. However, the longer a restaurant remains in the at-risk state, the greater its chances of failure. Following this idea, it is beneficial not only to consider whether a restaurant has been in an at-risk state, but also for how long. The latter is modeled by the variable-duration component in our VD-HMM.

We demonstrate the effectiveness of our model on the basis of 64,887 customer ratings drawn from 921 different restaurants in Phoenix, Arizona. Of these restaurants, nearly one quarter (24.43%) were classified as business failures, highlighting the importance of early warning mechanisms for management. We test different detection mechanisms using an out-of-sample setting, that is, whether one can detect the correct timing of failures for unseen businesses. To this end, the mean rating, a conventional metric from practice, achieves a balanced accuracy of around 50%, whereas conventional machine learning approaches would achieve 72.05%. Our model, in contrast, detects business failures with a balanced accuracy of 78.02%, a prediction that is even possible several months in advance. The model further identifies an at-risk state as being associated with a failure rate of 35.00%. In comparison, the failure rate in other states is lower by at least 41.41%. Hence, avoiding the at-risk state is associated with a considerable reduction in the estimated failure risk. The at-risk state is usually attained more than 78 weeks before failure of the business (i.e., exiting the market), giving marketers enough time to adjust their service portfolio.

This study has several implications with regard to both practice and research in marketing. In terms of practice, our model presents a data-driven early warning tool to prevent business failures by monitoring online customer ratings. Once a business arrived at the at-risk state, marketers would be able to take preventive actions in time, for example, by adjusting their service offering. Hence, these early warnings may help marketers meet their customers' demands and ultimately prevent their businesses from failing. To this end, customer ratings have obvious benefits. They are not only a direct measure of customer satisfaction but are also publicly available, thus allowing them to be leveraged in data-driven modeling. Regarding research, we develop a novel, variable-duration HMM that considers the duration of latent states (as opposed to only the latent states, as in a traditional HMM). The use of a VD-HMM is beneficial for all applications in which the latent dynamics are affected by repeated exposure. Based on our model, we show that customer ratings can be used to predict the risk of failure.

The remainder of this paper is organized as follows. In Section 2, we present related work on monitoring business performance via customer ratings. In Section 3, we develop a tailored, variable-duration HMM that models rating sequences in order to predict business failures. In Section 4, we describe the empirical setting, and in Section 5, we compare our model against common baselines from business decision making. In Section 6, we discuss findings and managerial implications.

2. Related Work

To infer insights concerning customers and markets, scholars in marketing frequently mine user-generated content (Fader and Winer 2012). For example, marketing scholars have analyzed user-generated content from social networks (e.g., Toubia and Stephen 2013, Wang et al. 2015). In the context of customer reviews, scholars applied text mining to derive not only information on market structure (Lee and Bradlow 2011, Netzer et al. 2012) but also on customer preferences (Archak et al. 2011, Culotta and Cutler 2016). Other scholars used sentiment analysis to study customer reviews and the relationship between customer opinions—positive or negative—and business performance (Sonnier et al. 2011, Tirunillai and Tellis 2012, Ludwig et al. 2013). Besides using qualitative data such as texts, scholars extensively used online customer ratings to measure business performance, as summarized in the following.

2.1. Relationship Between Online Ratings and Business Performance

Customer satisfaction has been considered an important dimension for many marketers when monitoring business performance, as it should theoretically precede sales (Ittner and Larcker 2003). Thus, scholars have extensively studied the relationship between customer satisfaction, reflected by online ratings (Ho et al. 2017, Schneider et al. 2021), and key financial figures such as revenue and sales.

Previous research on online ratings has repeatedly confirmed a positive relationship between ratings and business performance (see Appendix A in the e-companion for a literature summary). Ratings drive sales for a variety of products such as books (e.g., Chevalier and Mayzlin 2006) and movies (e.g., Dellarocas et al. 2007, Chintagunta et al. 2010). A meta-analysis of effect sizes is provided in Babić Rosario et al. (2016). In some works, the customer rating is augmented further by a score quantifying the textual sentiment of the written review (e.g., Archak et al. 2011). Prior research has primarily used an explanatory approach to study the correlation between ratings and sales (e.g., Chevalier and Mayzlin 2006, Liu 2006, Dellarocas et al. 2007, Chintagunta et al. 2010, Zhu and Zhang 2010, Moe and Trusov 2011), yet a few have used ratings to predict sales (Liu 2006, Dellarocas et al. 2007, Archak et al. 2011). Nonetheless, these studies differ fundamentally from our work (see Appendix A in the e-companion). Whereas theses papers have used ratings to study *sales*, we use online ratings to predict *business failures*.

2.2. Indicators of Business Failures

Long studied in prior marketing literature, business failures are relevant when monitoring performance so

that early warnings for managers can be provided (e.g., Sharma and Mahajan 1980). Additionally, business failures are relevant for other disciplines such as banking (Sarkar and Sriram 2001) and accounting (Altman 1968). Given that various disciplines study business failures in different contexts, several synonyms for business failure have emerged, including bankruptcy (Laitinen 1991), organization mortality (Swaminathan 1996), and organizational collapse (Argenti 1976). Likewise, definitions of business failure vary. Here, we refer to Pretorius (2009) for an overview of definitions, which range from "losses to creditors" (Lussier 1995, p. 9) to "deaths of entire firms" (Henderson 1999, p. 291). In this paper, we adopt the latter definition.

Prior literature on modeling business failures has mainly focused on financial data. These studies find that business failures are correlated with financial ratios (Beaver 1966, Olsen et al. 1983) and weighted financial ratios (z-scores; Altman 1968). Besides these correlation analyses, scholars have applied techniques from explanatory modeling such as logit/probit models (Ohlson 1980) and Bayesian models (Sarkar and Sriram 2001). However, these works focus extensively on financial variables, specifically financial ratios (Altman 1968). These financial ratios can successfully signal failure risk, for instance, in banking (Sarkar and Sriram 2001); however, in the service sector, other measures for business outcomes are more common, such as customer satisfaction.

Marketers could theoretically also use financial indicators to monitor customer satisfaction. However, financial indicators have an obvious caveat: they are only an indirect measurement of customer satisfaction. This is because "quality drive[s] consumer satisfaction, which in turn drives customer buying behavior, which in turn drives profits" (Ittner and Larcker 2003, p. 5). Hence, for marketers, financial indicators are considered "symptoms of failures" (Sharma and Mahajan 1980, p. 83). To measure business outcomes, marketers require performance indicators that directly reflect customer satisfaction, such as online ratings (e.g., Ho et al. 2017, Schneider et al. 2021). To this end, our work studies the predictive value of using online ratings to forecast the probability of business failures. However, business failure may not depend solely on customer satisfaction—that is, quality—but also on operational inputs. High quality may come at the cost of excessive investments. Hence, to link both, we next discuss the literature on the service-profit chain (SPC; Heskett et al. 1994, Kamakura et al. 2002).

2.3. Cost Structure of Service Businesses

In the context of service businesses, prior literature has developed a framework linking customer satisfaction to costs: the SPC framework (e.g., Heskett et al. 1994,

Kamakura et al. 2002, Mittal et al. 2005). The SPC framework suggests that business outcomes (Rust et al. 1995) depend not only on customer satisfaction but also on operational inputs/costs (Kamakura et al. 2002). Hence, a business can either improve satisfaction or reduce costs (or both, i.e., dual emphasis; see Mittal et al. 2005) to positively influence business outcomes (Mittal et al. 2005). For example, restaurants may achieve a high level of satisfaction but at the expense of high (service) costs, due to which they are unsustainable in the long-term. When predicting business outcomes, we must take into account both satisfaction—measured via ratings—and costs.

In the context of our specific research setting, that is, restaurants, we can identify several cost drivers. Here, it is common to distinguish between equipment and service costs (for instance, see the operational efficiency model in Kamakura et al. 2002). In the context of equipment costs, literature suggests the size (Yoon and Jang 2005; Camillo et al. 2008; Parsa et al. 2011a, b, 2015; Mun and Jang 2018) and the rent level (Bayou and Bennet 1992, Parsa et al. 2011a, Mun and Jang 2018). In the context of service costs, there are the "costs of goods sold" (Mittal et al. 2005, p. 546). During modeling, this is often reflected by the price level, which is found to be an important determinant for restaurant failure (Bayou and Bennet 1992, Camillo et al. 2008, Mun and Jang 2018). Another factor in the context of service costs is service work. The latter is typically modeled based on data from salaries or wages (Bayou and Bennet 1992, Camillo et al. 2008, Mun and Jang 2018). Here, the literature links between-city variation in salaries to restaurant failures. However, this is in contrast to our work, where we perform a within-city analysis. In a within-city setting, the overall wage level should remain fairly equal and, hence, is omitted from our model. Instead, we focus on the number of seats as a proxy of service personnel. Outside of the SPC framework, a key determinant of restaurant failures is location. This finding has been empirically backed in prior literature; for example, Parsa et al. (2011b) found that location (encoded via restaurant density) is strongly correlated with restaurant failure (correlation coefficient, 0.9919). Hence, location is considered in our model as another predictor. To this end, following the SPC framework, we consider both customer satisfaction and operational costs to predict business failures.

2.4. Inferences from User Behavior with Hidden Markov Models

HMMs represent a flexible class of time-series models with latent dynamics (Rabiner 1989, Netzer et al. 2008); that is, an observable time series (the so-called emissions) undergoes transitions between a discrete set of unobservable (i.e., latent) states. The relationship between both observations and latent states is assumed to be of stochastic nature, that is, observations are modeled as noisy realizations of a latent state via an emission probability. The latent states, in turn, change with a specific transition probability. Based on this, a sequence of latent states can be inferred. In practice, latent states are associated with certain interpretations (e.g., by naming them "at risk"), and reaching a state is often used as a signal to trigger a predefined management action (Netzer et al. 2008, Ding et al. 2015, Ascarza et al. 2018).

Because of their flexibility, HMMs have seen frequent application in marketing science, for example, in the context of customer churn (Ascarza and Hardie 2013, Ascarza et al. 2018), purchase intent modeling (Montgomery et al. 2004, Abhishek et al. 2012, Ding et al. 2015, Hatt and Feuerriegel 2020), targeting (Montoya et al. 2010), and customer relationship dynamics (Netzer et al. 2008, Schweidel et al. 2011, Zhang et al. 2014a). Further examples include unobservable competitive promotions (Moon et al. 2007) and cross-selling analyses (Li et al. 2011).

HMMs are based on the Markov property, according to which the transition probability is based merely on the single previous state (Rabiner 1989). In other words, the probability is independent of the previous state sequence. HMMs can be extended such that they consider the duration of the previous latent state. Note that this requires profound changes regarding the underlying estimation routine: simply incorporating the duration of latent states is not feasible as the duration is also latent and thus unknown. Instead, both latent states and latent state durations must be modeled jointly. This is subsumed under the wider class of hidden *semi-Markov* models (HSMMs; Murphy 2002, Yu 2016).

In hidden semi-Markov models, the latent state sequence is relaxed into a semichain. Here, different model variants exist (Murphy 2002, Yu 2016). One variant (called an explicit-duration HMM) models the renewal probability of latent states, so that the renewal probability depends on the latent state duration (Chiappa 2014). This alters the expected duration of a latent state and is thus beneficial for applications in which one state is of longer duration than others (e.g., when states capture deep engagement or cognitive absorption). A different variant (called a variable-duration HMM) models the transition probabilities, so that the transition probabilities depend on the latent state duration (Murphy 2002). This fulfills the need of our research, as we want to make inferences regarding a business's failure risk based on how long the business was exposed to an at-risk state. Previous applications of HSMMs can be found outside marketing science (e.g., for analyzing DNA sequences; see Barbu and Limnios 2008), whereas we add a specific HSMM for marketers.

3. Model Development

3.1. Overview

We model the sequence of online ratings in order to predict the risk of failure. To this end, we utilize the HMM-based framework for three reasons. First, it considers the sequence of rating events and thus yields a dynamic model. Second, it assumes that ratings are stochastically linked to a latent service performance for which ratings represent noisy observations and, by modeling this stochastically, the prediction performance is improved. Third, the HMM-based framework models latent states, which, in our case, are relevant to decision making: if we observe an at-risk state, this indicates that the business may fail in the long run. This state thus calls for managerial actions by decision makers in order to restore performance to a well-running state.

Our proposed model takes the observed sequence of rating events (i.e., the number of stars) for each business as input. It also considers their order and the time between two consecutive ratings.[1] Business failures are likely to reveal considerable between-business variation (e.g., a chain restaurant might be unlikely to close as it draws a customer base despite poor ratings). Hence, we consider various covariates describing the between-business heterogeneity as part of our input. We further account for unobserved heterogeneity through a random-effects specification (as in, e.g., Netzer et al. 2008, Schweidel et al. 2011). Eventually, the model recovers a sequence of latent states and, based on these, infers the probability of a business failure.

The intuition of our model is as follows. Businesses such as restaurants have different levels of service performance over time. For example, a restaurant might have a well-rated service performance, but its service performance might deteriorate after its kitchen chef left. In our model, we capture the service performance through latent states. One state denotes when a restaurant is at risk of a business failure. The longer a restaurant is now exposed to the at-risk state, the more likely it is to fail (e.g., because of word of mouth or because repeat customers have moved on to other restaurants). Analogously, the longer a restaurant is in a state with well-rated service performance, the less likely it will leave the state (e.g., because of a growing base of repeat customers). Hence, we use the duration of a restaurant being in an at-risk state for predicting business failures (and, analogously, the duration of a well-running state for predicting survival).

Traditional HMMs are based on the Markov property, whereby latent dynamics consider the previous latent state (e.g., Montgomery et al. 2004, Netzer et al. 2008, Ascarza and Hardie 2013, Ding et al. 2015, Ascarza et al. 2018), but not the duration of the latent state. However, the duration of latent states is likely to be of value in our work: the longer the exposure to an at-risk state, the higher the likelihood of a business failure. At the same time, it should also make it more likely that a business remains at risk. This is similar to other applications in marketing where renewal probabilities change with repeated use (Fader et al. 2018). Following this motivation, we develop a custom model, that is, a variable-duration HMM, where the duration of latent states is explicitly considered.

To make predictions, our model has two emission components linking the latent states to observable outcomes. On the one hand, we use an emission component to formalize the state–rating link. On the other hand, we introduce a secondary emission component—as in Ascarza and Hardie (2013)—to predict business failures from the latent dynamics. As a result, both ratings and failures are emissions from the same latent state and, thus, can be estimated jointly. The model finally recovers a sequence of latent states. The latent states allow for managerially relevant interpretation. As in Ascarza et al. (2018), one state is later named the at-risk state. Our results show that failures in the at-risk state are twice as likely as for the other states.

Upon deployment, the model takes a rating sequence as input and then outputs the failure probability for the given business. Here, we strive for accurate predictions regarding *which* businesses will fail and *when*. To achieve this, we build upon a prediction framework similar to that of Ding et al. (2015) or Montgomery et al. (2004).[2] The former, that is, *which* business fails, is addressed by evaluating the model based on out-of-sample businesses from the holdout set. For this, the model returns an output F_i (with $F_i = 1$ in the case of failure). The latter, that is, *when* it fails, is addressed by choosing an appropriate F_i during estimation. If the model is estimated with data from one-month-ahead failures, it predicts the one-month-ahead probability of failure; if it is estimated with two-month-ahead failures, it predicts the two-month-ahead failure probability, etc.

3.2. Specification of the Variable-Duration Hidden Markov Model

Our proposed VD-HMM consists of five components: (1) the observable rating sequence, (2) latent states, (3) the state–rating emission component, (4) the state–failure emission component as an additional emission in order to predict business failures, and (5) the variable-duration transition mechanism between states. The resulting VD-HMM is schematically illustrated in Figure 1. The five components are detailed in the following.

3.2.1. Observable Rating Sequence. The VD-HMM models a sequence of ratings across different businesses. The ratings are usually given by a star rating

Figure 1. Proposed Variable-Duration Hidden Markov Model
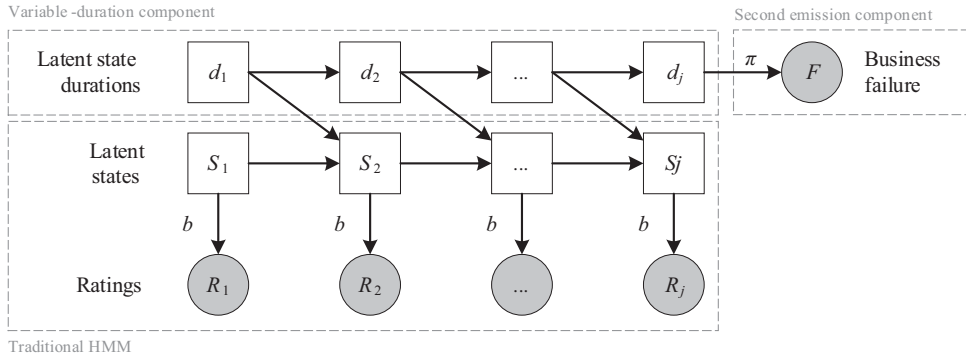
Notes. Shown is the VD-HMM for a given business *i*. Ratings represent the observable variables in the model, based on which failures are to be predicted. This is achieved by capturing latent dynamics behind service performance. Different from a traditional HMM, our model not only accommodates latent states but also their durations. The latter is responsible for the variable-duration component.

from a discrete set $\mathcal{R} = \{1, \ldots, R\}$. For instance, Yelp uses ratings from one to five stars. Other representations of ratings can be handled in a straightforward manner (shown later), simply by choosing an appropriate distribution in the emission. Formally, let $R_{ij} \in \mathcal{R}$ denote the rating score for business $i = 1, \ldots, N$ belonging to rating event $j = 1, \ldots, M_i$. Note that the rating sequences of each business can be different and, hence, the length of the sequence M_i is business specific. This is consistent with research in marketing science where a sequence of events, each associated with a time difference between two consecutive events, is modeled and where calendar-time effects are captured in a nonhomogeneous transition mechanism via a covariate denoting the time difference between two events (Montgomery et al. 2004, Ding et al. 2015).

3.2.2. Latent States. Each rating is associated with a latent state S_{ij}. Specifically, the VD-HMM assumes the existence of $\mathcal{S}$ different latent states, so that $S_{ij} \in \{1, \ldots, \mathcal{S}\}$ holds. These latent states are not directly observable; instead, each rating is linked stochastically to the latent states via the state–rating emission component. We later show how latent states can be recovered.

Latent states are also associated with a latent state duration $d_j(s)$ as follows. Let $d_j(s)$ denote the prior duration spent in a latent state s. Note that the variable $d_j(s)$ is—analogous to the latent state—also unobservable and thus latent. We model the latent state duration in discrete time, that is, in the number of ratings. This is consistent with research in marketing science where a sequence of events, each associated with a time difference between two consecutive events, is modeled (Montgomery et al. 2004, Ding et al. 2015). Modeling duration through the number of ratings has practical benefits. We expect that the failure risk increases when a large number of customers have been unsatisfied with the service experience. By using the number of ratings, we implicitly account for the

overall exposure (visitor frequency), which is advantageous for predicting the failure probability. For instance, an at-risk state might be even more dangerous for survival if the restaurant has only a few visitors, that is, is rated only rarely. By choosing our modeling approach for the latent state duration, we directly perform this adjustment before estimating the failure probability $\pi_i(j)$ (see Equation (3)).[3] For reasons of notation, let the $\mathcal{S}$-dimensional vector $d_j \in \{1 \ldots j\}^{\mathcal{S}}$ denote the latent state duration for all latent states $s = 1, \ldots, \mathcal{S}$. Again, we later discuss how the latent state duration can be recovered.

3.2.3. State–Rating Emission Component. The state–rating link introduces a so-called emission component $b^{(s)}(r)$. It defines a probability that a certain rating r from a business is observed given the current latent state s. Mathematically, the emission probability is written as

$$b^{(s)}(r) = P(R_{ij} = r \mid S_{ij} = s)$$
$$\text{for all } r \in \mathcal{R} \text{ and } s \in \{1, \ldots, \mathcal{S}\}. \quad (1)$$

The exact specification of $b^{(s)}(r)$ must consider that ratings are of a discrete nature. For this reason, we follow prior literature on ratings (e.g., Moe and Schweidel 2012, Lee et al. 2015) and model the emission probability as an ordered probit model, that is,

$$b^{(s)}(r) = P(R_{ij} = r \mid S_{ij} = s)$$
$$= \begin{cases} 1 - \Phi(\eta_s - c_1), & \text{if } r = 1, \\ \Phi(\eta_s - c_{r-1}) - \Phi(\eta_s - c_r), & \text{if } 1 < r < R, \\ \Phi(\eta_s - c_{R-1}), & \text{if } r = R, \end{cases} \quad (2)$$

where Φ denotes the cumulative standard normal distribution function, $c_1, \ldots, c_{R-1}$ are cut points, and η_s denotes state-specific intercepts for $s \in \{1, \ldots, \mathcal{S}\}$. Identifiability of the intercepts is ensured by setting η_1 to zero.

3.2.4. State–Failure Emission Component for Predicting Business Failures.
The VD-HMM has a second emission component in order to predict business failure after rating j. This is different from a traditional HMM, which is limited to only a single emission component.[4] This allows us to model both ratings and business failures as emissions from the same latent state, so that both can be estimated jointly. A similar idea has been put forward in the context of customer churns (Ascarza and Hardie 2013); however, we tailor it to our setting in the following. In our work, the secondary emission component is estimated with data from past business failures (F_i). Here, F_i is a binary variable that denotes whether a business has failed ($F_i = 1$) or whether the business is still in operation ($F_i = 0$). The exact meaning of the failure flag must be chosen during model estimation and allows one to calibrate the model to different prediction horizons: if the model is estimated with past data with F_i as one-month-ahead failures, it predicts one-month-ahead failures, etc. Hence, by varying which F_i is fed into the model during estimation, the prediction horizon can be chosen according to the needs of marketers.

Mathematically, the failure probability $\pi_i(j)$ for business i is modeled via logit and consists of three parts. First, it includes an intercept ω_0 that refers to the baseline probability. Second, the inference is based on the duration $d_j(s)$ in a latent state s. This is motivated by evidence that models in marketing benefit from considering duration times (e.g., Helsen and Schmittlein 1993, Schweidel et al. 2008). In our case, the intuition is that a longer duration in an at-risk state is linked to a higher failure probability. Put simply, we model the failure probability to be a function of the current state and how long the state has been in effect. For this reason, we draw upon the duration $d_j(s)$ for a specific state s but where the effect varies across states because of a state-specific coefficient ω_s. The duration is entered as a logarithmic value because it helps to better distinguish smaller durations. Third, we control for between-business variation. Altogether, this yields the following logit model:

$$\pi_i(j) = \text{logit}^{-1}\left[\underbrace{\omega_0}_{\text{baseline probability}} + \sum_{s=1}^{\mathcal{S}} \underbrace{\omega_s \log\left(d_j(s)\right)}_{\text{latent state duration}} \right.$$
$$\left. + \underbrace{\nu z_{ij}}_{\text{business covariates}} \right], \tag{3}$$

with coefficients $\omega_0, \omega_1, \ldots, \omega_{\mathcal{S}}$ and ν.

The failure probability $\pi_i(j)$ is then linked to the binary variable F_i (where $F_i = 1$ is used to indicate a failure and $F_i = 0$ otherwise). We model F_i to follow a Bernoulli

distribution. Hence, the corresponding likelihood is derived as

$$\log\left(P\left(F_i \mid R_{i1}, \ldots, R_{ij}\right)\right) = F_i \log\left(\pi_i(j)\right)$$
$$+ (F_i - 1)\log\left(1 - \pi_i(j)\right). \tag{4}$$

The variable F_i is later set during estimation according to the desired forecast horizon. One can predict failures one month ahead by estimating the model with historical data on F_i from one-month-ahead failures; one can predict failures two months ahead by estimating the model with historical data on F_i from two-month-ahead failures; etc.

3.2.5. Variable-Duration Transition Mechanism Between Latent States.
The transition component describes the probability of moving from a latent state s to a new state s'. In a traditional HMM, the probability of moving to a new state s depends *only* on the current state s', whereas, in contrast, our model considers the duration $d_j(s)$ of the latent state.

Two types of transitions can occur. These distinguish the new latent state s' given that the model exits state s: (1) A transition occurs where the latent state remains the same, that is, $s = s'$. We refer to this transition as *recurrent*. Such a self-transition occurs with a probability γ_s, as defined below. (2) The latent state can transition to a different state, that is, $s \neq s'$. We refer to this as *nonrecurrent*. Such a transition occurs with probability $1 - \gamma_s$. Both transitions—recurrent and nonrecurrent—are detailed in the following:

Recurrent part ($s = s'$). The recurrent part models the probability with which the current state is maintained. Hence, this allows us to control the duration of a latent state. This is done by modeling the (discrete-time) hazard function of staying in the current state. Formally, the probability of a self-transition is defined via a logit model

$$\gamma_s\left(d_j, \Delta_{j,j+1}, \sigma_{ij}\right) = \text{logit}^{-1}\left[\lambda_0^s + \lambda_1\log\left(1 + d_j(s)\right)\right.$$
$$\left. + \lambda_2\log\left(1 + \Delta_{j,j+1}\right) + \lambda_3\,\sigma_{ij}\right] \tag{5}$$

with additional variables as follows. The parameters λ_0^s, λ_1, λ_2, and λ_3 are estimated from the data. We point out that the intercept λ_0^s is state specific; that is, we yield different intercepts $\lambda_0^1, \ldots, \lambda_0^{\mathcal{S}}$ for each state. This models a different propensity for self-transition in certain states. As we shall see later, reaching an at-risk state is associated with a high probability that the business will remain in this state. The variable $d_j(s)$ is the latent state duration. It turns the model into a *variable-duration* HMM, so that certain states become more sticky because of repeated exposure. In other words, the probability of self-transition depends on the prior duration in a state. The variable $\Delta_{j,j+1}$

denotes the time difference (in calendar time) between two consecutive ratings. We intentionally include $\Delta_{j,j+1}$ as it allows us to control for the elapsed time between the two ratings. Therefore, the parameter λ_2 captures the effect of calendar time. This is consistent with research in marketing science (Montgomery et al. 2004, Ding et al. 2015), where, analogously, the time difference between two events is modeled via a nonhomogeneous transition mechanism. By including $\Delta_{j,j+1}$, we implicitly consider the popularity of a business. The more ratings a business receives, the shorter the time difference $\Delta_{j,j+1}$ should be. Finally, the variable σ_{ij} controls for the review sentiment.

Nonrecurrent part ($s \neq s'$). The nonrecurrent part models the transitions from a latent state s to a different latent state s' with $s' \neq s$. The transition from s to a different s' is modeled by a separate probability $\Gamma_{ss'}$. This yields a matrix Γ for which the rows are stochastic; that is, $\Gamma_{ss'} \in [0,1]$ and $\sum_{s'=1}^{\mathcal{S}} \Gamma_{ss'} = 1$ for all $s, s' \in \{1, \ldots, \mathcal{S}\}$. Recall that after exiting state s, all potential subsequent states of a nonrecurrent transition are given by $s' \in \{1, \ldots, \mathcal{S}\} \setminus \{s\}$. By definition, self-transitions are already modeled via the recurrent part, and, hence, the diagonal elements of Γ are set to zero.

The transition component then combines both the recurrent and nonrecurrent parts. It thus yields a transition probability

$$P\left(S_{i,j+1} = s' \mid S_{ij} = s, d_j\right)$$
$$= \begin{cases} \gamma_s(d_j, \Delta_{j,j+1}, \sigma_{ij}), & \text{if } s' = s, \\ (1 - \gamma_s(d_j, \Delta_{j,j+1}, \sigma)) \times \Gamma_{ss'}, & \text{if } s' \neq s. \end{cases} \quad (6)$$

We also experimented with alternative specifications, yet with inferior results.[5]

3.3. Relationship to Traditional HMMs
The VD-HMM has two clear differences in comparison with a traditional HMM. First, the VD-HMM introduces a variable-duration component, so that transitions depend not only on the previous latent state but also on the duration of being in a latent state. Here, we note that the traditional HMM is a special case of our VD-HMM. In fact, one can yield a traditional HMM with a stationary transition matrix by fixing λ_1, λ_2, and λ_3 all at zero. Second, our model relaxes the assumption of homogeneous transitions inherent to traditional HMMs. In contrast, the transitions in the recurrent part of our model are time dependent, as they consider the time difference $\Delta_{j,j+1}$ (in calendar time). Hence, the evolution of states is nonhomogeneous.

The VD-HMM is a special case of a broader class of so-called hidden semi-Markov models (Yu 2016). These relax the assumption of traditional HMMs, namely, that the so-called dwell time of a Markov chain—that is, the number of rating events a chain stays in a given state—follows a geometric distribution. Hence, the most likely dwell time for a given state amounts a priori to one in a traditional HMM. Hidden semi-Markov models allow for more flexible dwell time distributions.

3.4. Model Estimation
We estimate all model parameters using a Bayesian framework. Specifically, we derive the log-likelihood of the VD-HMM and then apply Markov chain Monte Carlo so that the model parameters are directly sampled from their posterior distribution; see Appendix B in the e-companion for details.

To account for unobserved heterogeneity across businesses, we report parameter estimates from a random-effects model. For this, we leverage a hierarchical Bayes procedure through the use of appropriate priors. Specifically, we adopt a random-effects specification as in Netzer et al. (2008) and Schweidel et al. (2011); that is, intercepts in the transitions (λ_0^s) are allowed to vary across businesses. We do not use a full random-effects model as we perform predictions on out-of-sample (i.e., unseen) businesses for the same reasons as in Schweidel et al. (2011).

The number of latent states, $\mathcal{S}$, is determined analogously to prior research (Netzer et al. 2008, Ascarza and Hardie 2013, Ascarza et al. 2018). Accordingly, we estimate a series of VD-HMMs with a varying number of latent states $\mathcal{S}$. Then, we selected the model with the best fit. In our experiments, we vary the number of latent states between $\mathcal{S} = 1$ and $\mathcal{S} = 4$. Setting $\mathcal{S} = 1$ yields a baseline with no latent structure. This model is equivalent to a logistic regression model with business covariates and a factor controlling for the number of ratings (note that $d_{M_i}(1) = M_i$ if $\mathcal{S} = 1$). Setting $\mathcal{S}$ to five or larger is not reasonable, because ratings are reported on a one-to-five scale and we thus yield more latent states than observations.

Model selection is based on the prediction performance of the model. This is analogous to prior HMM-based research (Sismeiro and Bucklin 2004) and also advocated in the literature on Bayesian modeling (Gelman et al. 2014). We explicitly refrain from using information criteria for the reason that they judge the overall model fit. They thus put emphasis on the state–rating emissions because of their larger parameter space, while the state–failure emission suffers from overfitting. However, our objective is an accurate forecast of business failures. Therefore, we perform model selection based merely on the prediction performance with regard to business failures. As is common in predictive modeling, performance is measured by the area under the curve (AUC) from the receiver operating characteristic (ROC) curve.

3.5. Recovering the Latent State Sequence

We describe how we recover the latent states from a given rating sequence. Given a rating sequence $R_{i1}, \ldots, R_{ij}$, the objective is to determine the most likely latent state sequence $s_1^*, \ldots, s_j^*$. It is defined by

$$s_1^*, \ldots, s_j^* = \underset{s_1, \ldots, s_j \in \mathcal{S}^j}{\arg \max} \, P\Big(S_{i1} = s_1, \ldots, S_{ij} = s_j \mid R_{i1}, \ldots, R_{ij}\Big). \quad (7)$$

The computation is based on the Viterbi algorithm (Rabiner 1989), which is carefully adapted to our VD-HMM; that is, it is tailored to the variable-duration transition mechanism inside the VD-HMM.

4. Setting
4.1. Data Set

We obtained our rating data from Yelp, which has been used in prior research (e.g. Luca and Zervas 2016). Our sample consists of all ratings between January 2010 and December 2017 for restaurants listed in Phoenix, Arizona.[6] This amounts to 934 distinct restaurants. We manually obtained ground-truth labels concerning *which* restaurants experienced a failure and *when* (i.e., the closing date). We followed the definition in (Henderson 1999, p. 291) whereby restaurant closings due to relocation, lease renewal, or retirement were not labeled as instances of failure.[7] We removed 13 restaurants that had closed by the end of the observation period but for which the reasons were unclear. This left us with a sample of 64,887 ratings from 921 restaurants, with an overall failure rate of 24.43% (225 restaurants).

We build upon a prediction framework similar to that used in Ding et al. (2015) and Montgomery et al. (2004).[8] This allows us to test our model's ability to predict which and when restaurants fail. As is common in machine learning (Murphy 2012), we randomly split the 921 restaurants into three sets as follows: (1) The *training set* has 500 restaurants (54.29%). We used it solely for estimating the model parameters. (2) The *calibration set* (called "validation set" in machine learning) has 100 restaurants. We used it to find optimal cutoffs (thresholds) for the VD-HMM as well as the baseline models, so that the failure probability is mapped onto a binary prediction.[9] (3) The *test set* has 321 restaurants. We used it to measure the out-of-sample performance, that is, how well the model generalizes to a hold-out sample of previously unseen restaurants. If not stated otherwise, the performance is reported out-of-sample, that is, from the test set.

In our evaluation, we later compare different prediction horizons in which we evaluate whether restaurant failures can be predicted one month ahead, two months ahead, etc. For each, we used a different binary variable F_i denoting a failure in the given time horizon (i.e., $F_i^{1\text{-month}}$, $F_i^{2\text{-month}}$, etc.) and, based on this, evaluated the model. The variables $F_i^{1\text{-month}}$, $F_i^{2\text{-month}}$, etc., are computed based on the ground-truth data regarding which and when restaurants closed because of failure. Crucially, the model is not retrained for this task, but the cutoff δ is calibrated for a given prediction horizon using the calibration sample. This yields potentially different cutoffs for predicting the one-month-ahead, two-month-ahead, etc., failure risk.

4.2. Variable Description

We collected various business-specific variables for each restaurant in order to control for the between-business heterogeneity (see Table 1). In accordance with earlier rating-related research (e.g., Chen and Lurie 2013), our choices are as follows: (1) *restaurant density*, measured as the number of other restaurants within a 500-meter radius; (2) *restaurant age*, measured in months; (3) *check-in rate*, defined as the total number of check-ins divided by time horizon; (4) *chain status*, given by a binary variable that equals one if a restaurant belongs to a chain with more than 14 outlets nationwide (Luca and Zervas 2016); and (5) *restaurant category*, encoded based on the same classification used by Yelp. The most popular restaurant categories in our data set are American (27.73%), Mexican (17.48%), and Asian (12.49%). Because each restaurant

Table 1. Model Variables and Summary Statistics

Variable	Mean	SD
Panel A: Model variables at rating level		
Rating valence R_{ij} (discrete, 1–5 stars)	3.70	1.46
Time lag $\Delta_{j,j+1}$ between ratings (in days)	19.39	43.82
Review sentiment σ_{ij}	0.08	0.37
Panel B: Covariates $Z_{I,T}$ at the restaurant level		
Restaurant density	12.71	14.81
Restaurant age (in months)	48.10	29.48
Check-in rate	5.60	6.66
Chain status (affiliated = 1; independent = 0)	0.29	0.45
Restaurant categories (true = 1; false = 0)		
American	0.26	0.44
Asian	0.12	0.33
Cafe	0.09	0.28
Fast Food	0.11	0.31
Mexican	0.17	0.38
Pizza	0.09	0.29
Salad	0.03	0.19
Specialty food	0.05	0.23
Restaurant size (in square meters)	228.81	259.85
Rent level (Zillow Rent Index)	1492.52	179.94
Price level ($–$$$$) (true = 1; false = 0)		
$ = $10 or under	0.54	0.49
$$ = 11–30	0.45	0.49
$$$ = 31–60	0.01	0.09
$$$$ = over $61	0.00	0.00
Number of seats	98.09	103.88

Note. SD, Standard deviation.

potentially belongs to multiple categories predefined by Yelp, we assigned each restaurant to the most popular (as measured by numbers of restaurants on Yelp belonging to that category).

Informed by prior literature (e.g., the SPC framework), we additionally include variables that should reflect the cost structure of restaurants. These are (6) *restaurant size* (in square meters), (7) *rent level* (i.e., the Zillow Rent Index, analogous to Barron et al. 2021), (8) *price level* (according to Yelp classification into $–$$$$), and (9) *number of seats* as a proxy for service work.

In accordance with previous studies (Archak et al. 2011), we measured the textual sentiment of reviews as follows. Let N_{pos} denote the number of positive words in a review and N_{neg} the number of negative ones. Furthermore, let N_{total} refer to the total number of words in that review. Then the ratio $(N_{pos} - N_{neg})/N_{total} \in [-1, +1]$ gives the review sentiment σ_{ij}. Here, we followed common practice in sentiment analysis. This means that we subjected the text to additional preprocessing steps (Berger et al. 2020) that removed the punctuation and discarded stop words and preformed stemming. We tested the robustness with a variety of dictionaries—the General Inquirer dictionary as shipped in the Harvard IV software, the QDAP dictionary, and the AFINN dictionary—leading to conclusive findings. In the following, results from the latter dictionary are reported, as it is particularly tailored to colloquial expressions in user-generated content.

4.3. Comparison of Rating Distribution Across Open vs. Closed Restaurants
Aggregated rating metrics, such as the mean, variance, or entropy, hardly differ between failed and open restaurants (see Figure 2). For example, the median rating amounts to 3.65 for both. The similarity is further underlined by a Wilcoxon–Mann–Whitney test comparing the mean ratings across open and

closed restaurants. The test returns a *p*-value of 0.91, thus pointing to equally distributed mean ratings. Therefore, aggregated rating metrics are unlikely to entail predictive power with regard to business failures. This explains why the above rating metrics add little prediction power and motivate our model, in which the complete rating sequence is considered.

4.4. Baselines
In our experiments, we compare our VD-HMM against a traditional HMM (Rabiner 1989) that is tailored to our setting. To this end, the HMM includes a secondary emission component as in our VD-HMM, which allows the HMM to predict failures. The HMM has access to the same data as our VD-HMM. Specifically, the HMM uses the same business covariates in the state–failure emission component for making predictions. Different from our VD-HMM, it models only latent states but without considering latent state durations and, therefore, assumes a stationary transition matrix.

Furthermore, we present a series of classification models from machine learning that serve as our baselines. Analogous to our VD-HMM, these models make dynamic predictions of failures F_i. However, as a difference, the baseline models rely upon feature engineering rather than modeling the actual sequence of ratings. Specifically, we experiment with different combinations of features and prediction models as follows.

Our choice of features is informed by prior literature (Godes and Mayzlin 2004, Dellarocas and Narayan 2006, Moe and Trusov 2011) and includes the different variants: (1) we consider only the average rating of each restaurant; (2) we use different rating metrics ζ_{ij}, namely, the mean rating, the volume of ratings, and the volatility as measured by statistical entropy (referred to as simple rating dimensions); (3) we expand previous set of rating metrics (named

Figure 2. Rating Distribution at the End of the Observation Period Across Open and Closed Businesses
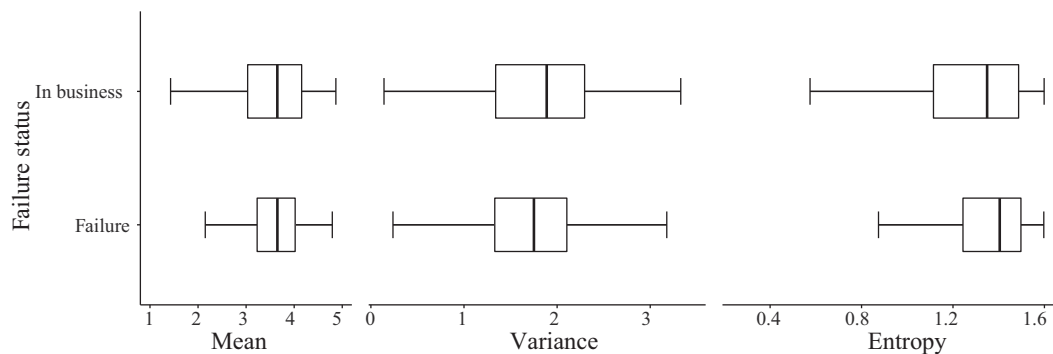
Notes. The plots compare the distributions of ratings for restaurants that had experienced a failure and those that were still in business by the end of our observation period. The similarity between the distributions explains the weakness of summary statistics of raw ratings as predictors of business failure and the motivation for the modeling approach adopted in this study.

advanced rating metrics) by additionally including the coefficient of variation and the distribution of ratings (i.e., the share of one-stars ratings, the share of two-stars ratings, etc.); (4) we draw upon the same business-specific covariates z_{ij} as in our VD-HMMs; and (5) we combine both business covariates z_{ij} and advanced rating metrics. We note that all features are dynamic; that is, for a prediction at observation j, only the first j ratings are considered for computing ζ_{ij}. We also experimented with the last $n = 5$ ratings instead of using the mean as part of our robustness checks; this resulted in similar conclusions.

The above features are inserted into different prediction models, namely, a logistic regression model and a random forest. The former is modeled analogously to our state–failure emission from Equation (3), so that the same time-dependent covariates z_{ij} are considered. For instance, for rating metrics ζ_{ij} and covariates z_{ij}, the failure risk $\pi_i(j)$ is predicted via

$$\pi_i(j) = \text{logit}^{-1}\left[\underbrace{\varphi_0}_{\text{baseline probability}} + \underbrace{\varphi_\zeta \, \zeta_{ij}}_{\text{rating metrics}} + \underbrace{\nu \, z_{ij}}_{\text{business covariates}} \right] \quad (9)$$

with coefficients φ_0, φ_ζ, and ν. The second baseline model, the random forest, is known to be a powerful machine learning classifier that can adapt to various degrees of nonlinearity (Murphy 2012). Both baseline models are dynamic in the sense that they depend on features calculated for rating j.

The baselines were estimated in the same dynamic manner as in our VD-HMMs. All hyperparameters, such as the number of decision trees to be grown and the number of predictors considered at each split, were tuned via grid search and 10-fold cross-validation. Consistent with our approach to model selection, the baseline models were also optimized against the AUC during training. Altogether, this allows us to later establish that, because of the challenging nature of our prediction task, feature-based approaches are subpar; instead, one should model the actual rating sequence. Estimation results for the logit model are provided in Appendix C of the e-companion.

5. Empirical Results

As part of our model selection, we first determine the preferred model specification. For the selected model, we then provide an interpretation of the latent states, discuss its predictive power, and compare it with alternative classifiers.

Table 2. Overall Prediction Performance

Model variant	Latent states $\mathcal{S}$	In-sample AUC (training set)	Out-of-sample AUC (test set)
HMM	1	78.06	76.68
	2	78.57	78.50
	3	81.84	83.42
	4	82.12	82.68
VD-HMM	1	78.06	76.68
	2	78.69	77.05
	3	**83.94**	**85.20**
	4	83.39	84.52

Notes. In-sample (i.e., on the training set) and out-of-sample (i.e., on the test set) AUC values are in percentages. Best values are indicated in bold.

5.1. Overall Prediction Performance

In the following, we compare the performance of the VD-HMM across different numbers of latent states against a traditional HMM (see Table 2). We follow prior research (Sismeiro and Bucklin 2004) and report the out-of-sample AUC, that is, the ability to predict business failures on the test set. The best performance is obtained by the VD-HMM with three latent states. It attains an out-of-sample AUC of 85.20%. For comparison, a VD-HMM with two latent states or no latent dynamics appears inferior. Furthermore, our VD-HMM increases the out-of-sample AUC on the test set compared with the best HMM by 1.78 percentage points. Recall that the HMM has access to the same data as the VD-HMM, specifically, the same covariates in the state–failure emission. Altogether, the findings have two implications: first, including latent dynamics with three states is beneficial. Second, the variable-duration component helps in making better predictions. Given these outcomes, all subsequent analyses build upon the VD-HMM with three latent states.

5.2. Estimation Results
5.2.1. Characterizing the Latent States.
In the following, we examine the nature of the identified states. We find that one of the three latent states is linked to an increased failure risk and, hence, label it as the at-risk state. The other two states have a considerably lower failure risk but with different average rating scores (i.e., very high or very low ratings). Hence, we named them "well running" and "bad ratings but running."

The state-dependent rating distribution reveals considerable differences. Table 3 examines the posterior emission probabilities, that is, the probability distribution of rating scores for each of the latent states. We make the following observations:

• The well-running state has mostly positive ratings. The probability of observing a four- or five-star rating amounts to 88.55%. This results in a mean rating of 4.51 stars. Hence, these restaurants are associated with a large propensity to remain in business.

Table 3. State-Dependent Rating Distribution (in %)

| Rating valence | State | | |
	Well running	Bad ratings but running	At risk
5 (positive)	66.09 [64.79, 67.45]	32.31 [30.86, 33.81]	13.14 [11.65, 14.70]
4	18.75 [18.07, 19.45]	27.12 [26.34, 27.81]	15.08 [14.29, 15.92]
3	6.92 [6.56, 7.28]	14.69 [14.13, 15.21]	11.59 [10.97, 12.21]
2	4.62 [4.33, 4.92]	12.52 [12.01, 13.09]	14.26 [13.54, 14.97]
1 (negative)	3.62 [3.20, 4.07]	13.36 [12.50, 14.30]	45.93 [43.79, 48.09]

Note. Posterior emission probabilities (means) are shown, with 95% confidence intervals in brackets.

- The bad-ratings-but-running state emits mostly negative ratings. One- or two-star ratings appear with a combined probability of 65.71%. Accordingly, the mean rating amounts to 2.16 and is thus the lowest of all states. Nevertheless, these restaurants have a large propensity to remain in business.
- The at-risk state reveals a dispersed rating behavior. It has an almost uniform distribution of ratings. This is also reflected in a mean rating of 3.53, which ranges in between the other states.

Overall, we find that restaurants in the bad-ratings-but-running state are subject to a lower average rating than those in the well-running state. This motivated our naming. The state-dependent rating behavior also points to the challenges of our prediction task: states with a low risk entail either mostly positive or mostly negative ratings—yet such ratings are also likely to occur in the at-risk state. Our model can nevertheless discern the different states by modeling the complete rating sequence.

Despite negative ratings, restaurants in the bad-ratings-but-running state are largely secure against business failures. Luca (2011) offers an explanation: some restaurants are less dependent on ratings, as they have already built a customer base (e.g., because of a brand name or being older). For instance, the mean age of restaurants in the bad-ratings-but-running state is 4.92 years, compared with 3.69 years for the average restaurant in our sample. Furthermore, chains often belong to the bad-ratings-but-running state and remain in business even if poorly rated.

5.2.2. State–Failure Relationship. The state–failure emission component is responsible for making the predictions. It links the latent state duration and additional business covariates to the probability of business failure. The estimation results are displayed in

Table 4. Relationship Between Latent States and Business Failure

| Parameter | Posterior mean | 95% CI | |
		Lower	Upper
Intercept	−1.8205***	−2.1829	−1.5051
Latent state durations ωs			
Well-running state	−0.3239	−0.7166	0.0348
At-risk state	0.8862***	0.5447	1.2519
Bad-ratings-but-running state	0.2559	−0.1335	0.6285
Business covariates			
Restaurant age	−0.0008***	−0.0012	−0.0004
Restaurant density	−0.0035	−0.0191	0.0139
Check-in rate	−0.0304	−0.0728	0.0132
Chain status	−2.4051***	−3.3451	−1.5133
Restaurant size	−0.0015	−0.0045	−0.0016
Rent level	0.0016***	0.0001	0.0030
Price level	−0.2767	−0.8529	0.2937
Number of seats	0.0033	−0.0021	0.0094
Restaurant categories	✓		

Note. Posterior means are shown.
***$p < 0.001$.

Table 4. We find a positive relationship for the at-risk state. A longer duration in this state is associated with a higher failure risk. We observe an opposite relationship for the well-running and bad-ratings-but-running states. Each additional rating event corresponds to a lower failure risk. Altogether, the different effect on the failure probability explains the naming of the at-risk state. In order to yield an intuitive interpretation, we report the odds ratio: in keeping all covariates at their mean, a one standard deviation longer duration of the at-risk state increases the odds ratio of a failure from 0.16 to 0.40. Put simply, this corresponds to an increase in the failure probability of 14.35 percentage points, amounting to 28.41%.

Table 4 also reports the coefficients belonging to the business covariates. These are supposed to control for between-business heterogeneity and yield results in line with our expectations. For example, failure risk is negatively linked to age: the older a restaurant, the lower its risk of failure. The coefficient for restaurant density is small and not significant, suggesting that this variable offers little or no predictive power. Chains exhibit a lower failure risk overall than independent restaurants. A possible reason is that chains may be able to cross-subsidize poorly running sites.

5.2.3. State-Specific Failure Rates. In what follows, we compare the failure rates of the different latent states. Mathematically, we first recovered the latent states S_{i,M_i} from the observable ratings and then compared their distribution across restaurants still in business and those classified as failures. This distribution is given in Table 5. In the at-risk state, 35.00% of the restaurants fail. In contrast, the failure rates for the

Table 5. Distribution of Recovered Latent States (at End of Observation Period)

Status	Total	Frequency of recovered latent state		
		Well running	At risk	Bad ratings but running
In business	696	338	130	228
Business failure	225	80	70	75
Failure rate (in %)	24.43	19.14	35.00	24.75

other states are considerably lower. These amount to a mere 19.14% (well running) and 24.75% (bad ratings but running). These findings highlight again the strong link between the at-risk state and business failures.

5.2.4. Transitions Dynamics. The evolution of the latent state sequence is described by the transition dynamics. In the proposed VD-HMM, the transition mechanism consists of two parts, namely, a recurrent and a nonrecurrent part. The recurrent part models self-transitions, that is, the probability of remaining in the current state. The nonrecurrent part denotes the probability of transitioning to a different state. Both are studied in the following.

5.2.4.1. Recurrent Transition Dynamics. The recurrent part specifies self-transitions where the restaurant remains in the current state. It is given by a function that depends on (1) the latent state duration, as measured by λ_1; (2) the time elapsed since the last rating, as measured by λ_2; and (3) the sentiment of the most recent review, as measured by λ_3. With regard to the former, we find that the posterior mean of λ_1 amounts to −0.04. Hence, self-transitions are associated negatively with the latent state duration. Put simply, a longer exposure to a state makes it more likely that the restaurant will leave this state. With regard to the second element, we observe that the posterior mean of λ_2 is −0.20. Hence, larger delays between individual ratings make it more likely that the restaurant will have switched to a different state. Furthermore, we find that the posterior mean of λ_3 equals −0.01. Hence, a positive review sentiment decreases the likelihood of a recurrent transition but only to a slight extent. Furthermore, we find that the mean of the business-individual effects, and thus the propensity for self-transition, differs across states. This relationship is visualized in Figure 3, which compares the effects from the latent state duration (gray/black color) and the time since the last rating (x-axis). As may be seen, the probability for self-transition is generally highest for the well-running state and lowest for the bad-ratings-but-running state. In both cases, it diminishes with a larger time since the last rating.

5.2.4.2. Nonrecurrent Transition Dynamics. Table 6 reports the transition probabilities of changing to a different state. Most changes from the well-running state are to the bad-ratings-but-running state. The bad-ratings-but-running state reverts with 81.10% probability back into the well-running state. However, some restaurants (18.90% probability) also moved from the bad-ratings-but-running state to the at-risk state.

5.3. Predictive Power for Business Failures
5.3.1. Out-of-Sample Comparison. Next, we examine the ability of the VD-HMM in predicting restaurant failures. Table 7 compares our VD-HMM against a series of baseline models. These use machine learning to make dynamic predictions. Furthermore, they use same data, but operate on features rather than modeling the complete rating sequence. All comparisons are made for the out-of-sample AUC, that is, the prediction performance as measured on the test set. Overall, we find that our proposed VD-HMM is superior by a considerable margin. Indeed, the VD-HMM bolsters the AUC from the best machine learning model by 6.24 percentage points and the balanced accuracy by 5.93 percentage points. Both improvements are statistically significant at the 0.1% level. Overall, the VD-HMM achieves an AUC of 85.20% and a balanced accuracy of 78.02%.

The baseline, which relies solely on the mean rating, scored poorly, with the AUC being close to a random guess (i.e., 50%). This suggests that the mean rating lacks predictive power. In other words, it is ineffective for service marketers to monitor the mean rating when drawing inferences as in this study. This is expected given that the descriptive statistics did not reveal any differences in the mean ratings among open and closed restaurants.

We further make the following observations. First, including more predictors in the feature-based models improves the performance. Second, we obtain considerable performance improvements when modeling the complete sequence of customer ratings. When comparing the baseline models with business covariates against the VD-HMM, the relative gain in the AUC by the latter amounts to 11.10 percentage points. This improvement must be solely attributed to the predictive power of ratings. Third, the performance of the dynamic logit and the random forest are largely on par. Fourth, a closer examination of the confusion matrix reveals that the VD-HMM competes well in

Figure 3. Dynamics of Self-Transitions
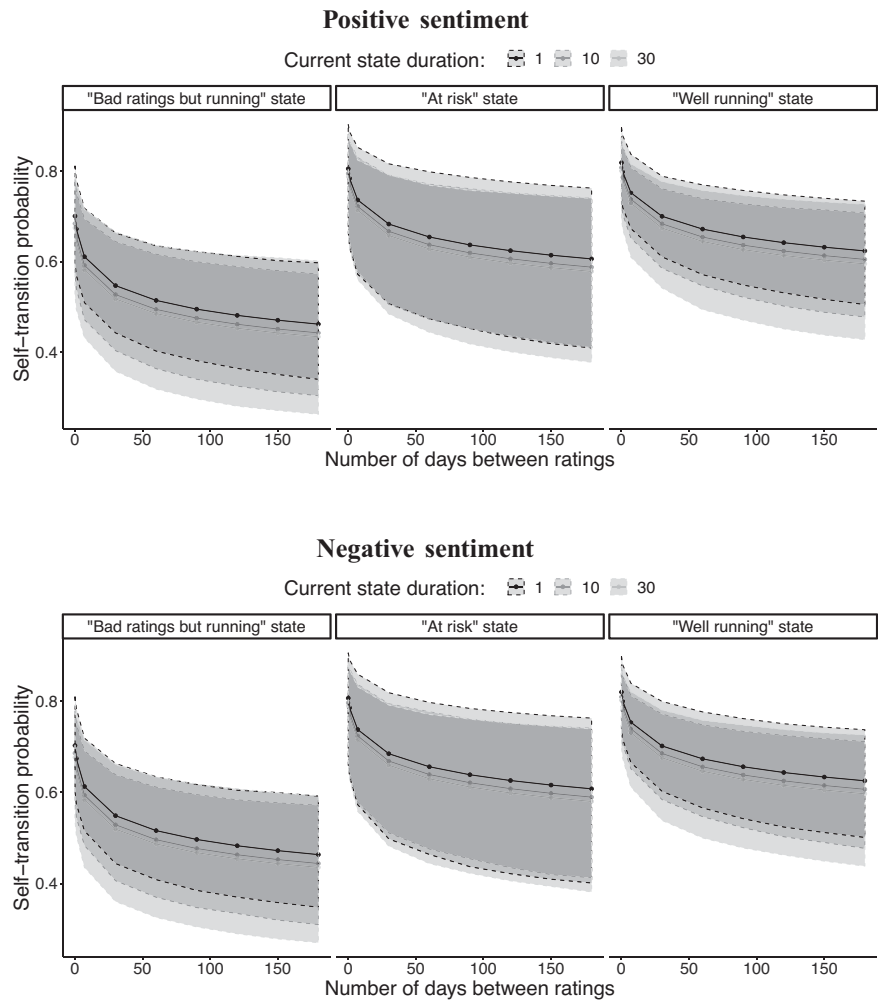
Notes. The recurrent part models self-transitions, that is, the probability of remaining in the same state. It depends on both the latent state duration (shown by differently shaded lines), the time since the last rating (shown on the *x*-axis), and the review sentiment σ_{ij} (top panel, positive sentiment; bottom panel, negative sentiment). Longer exposure to a state decreases the likelihood that the state will be maintained. Larger delays between ratings increase the propensity to switch to a different state. Results are shown for the average business-specific effect. The shaded area refers to the 95% CI.

detecting both restaurants that experienced a failure (i.e., sensitivity) and those that remained in business (i.e., specificity). This sensitivity–specificity trade-off is also reflected in the F1 score. The difference in terms of the F1 score between the best machine learning model and our VD-HMM amounts to 15.83%. Fifth, the VD-HMM is superior to the HMM, despite having access to exactly the same data. Hence, the performance improvement is attributed exclusively to the model formulation that appears to be of better fit.

5.3.2. Sensitivity of Prediction Performance Across Restaurant Subgroups. Table 8 conducts a sensitivity analysis across different subgroups of restaurants. For instance, the performance with regard to predicting business failures among restaurants of different ages is fairly similar. The prediction performance is lower

for restaurants affiliated with chains (AUC of 70.34%) compared with independent restaurants (AUC of 82.47%); however, this can be attributed to the small number of business failures among chains (i.e., only 5 out of 120 restaurants from chains experienced a failure). Altogether, our sensitivity analysis suggests that there is little variability, as the prediction performance remains robust.

5.3.3. Predicting Business Failures Early in Advance. An important question for business decision making concerns how far in advance the failure of a restaurant can be predicted. Table 9 compares the performance when making such a prediction several months in advance. Again, the VD-HMM attains the best performance. For example, choosing a forecast horizon of six months yields an AUC of 82.70%. When compared

Table 6. Propensity of Transitioning to a Different State

Current state	Next state		
	Well running	At risk	Bad ratings but running
Well running	—	44.05 [31.98, 56.81]	55.95 [43.19, 68.02]
At risk	66.97 [52.75, 80.48]	—	33.03 [19.52, 47.25]
Bad ratings but running	81.10 [65.44, 95.89]	18.90 [4.11, 34.56]	—

Notes. The table shows posterior means with 95% confidence intervals in brackets. The nonrecurrent part of the transition component specifies the probability of moving from the current state to a different one. Therefore, all self-transitions on the diagonal are omitted and are discussed as part of the recurrent part.

with the best machine learning baseline, it is also better at a statistically significant level. Notably, simply looking at the mean rating remains ineffective: the mean rating hardly surpasses a random guess with an AUC of around 50%, and, hence, predictive power is lacking. However, if appropriately modeled, ratings encode valuable information that facilitates long-term forecasts. Reiteratively, the improvement of the VD-HMM over the HMM is statistically significant for most forecast horizons.

5.3.4. Interpreting the At-Risk State as a Leading Indicator of Business Failures. The at-risk state is particularly relevant to business decision makers as it is associated with an elevated risk of business failure. However, it represents a valuable tool for decision makers only if they have sufficient time to implement remedial actions. Hence, we now study how far in advance the at-risk state is attained by a restaurant before its failure date. For this, we base our analysis on the actual timings of when restaurants closed. We then use the Viterbi algorithm to recover the latent state sequence from each restaurant. We further filter for all restaurants subject to failure. Afterward, we compute how long the at-risk state had already existed prior to the closing date. The corresponding lead time is reported in Table 10.

As we see, half the restaurants attained the at-risk state at least 78.43 weeks beforehand. One quarter of the restaurants arrived at this state more than 136.65 weeks in advance. Hence, the at-risk state should be interpreted by business decision makers as a leading indicator of elevated failure risk and thus seen as providing an early warning in due course, thus allowing for the implementation of appropriate managerial action.

Based on the above, one can further observe that the duration spent in the at-risk state is also important. Put simply, how long a restaurant has been in the at-risk state is linked to its likelihood of experiencing a business failure. Because of this, restaurants should be particularly worried about being in the at-risk state for an extended duration. Very few restaurants failed after being in the at-risk state only for a short duration, whereas 75% of the failed restaurants had been in the at-risk state for a long duration (i.e., for at least 31.96 weeks). This pattern is explained by the above estimation results (ω_s), implying that each additional time period spent in the at-risk state is linked to a larger failure probability.

6. Discussion
6.1. Summary of Findings
In this paper, we study the efficacy of predicting restaurant failures from online ratings. Therefore, we

Table 7. Comparison of Prediction Performance

Model	Features	AUC	Balanced accuracy	F1 score	Specificity	Sensitivity
Dynamic logit	Mean rating	51.60	47.13	30.96	50.21	44.05
Random forest	Mean rating	51.60	49.44	27.96	67.93	30.95
Dynamic logit	Rating dimensions	64.47	58.14	41.78	60.34	55.95
Random forest	Rating dimensions	56.45	55.69	39.30	57.81	53.57
Dynamic logit	Business covariates	72.85	64.52	48.93	61.18	67.86
Random forest	Business covariates	73.32	66.26	50.51	73.00	59.52
Dynamic logit	All	79.09	70.25	55.45	73.84	66.67
Random forest	All	78.64	68.04	52.83	86.08	50.00
Proposed VD-HMM		**85.22**	**79.84**	**70.24**	**89.45**	**70.24**

Notes. The table shows out-of-sample performance (i.e., based on test set) in percentages assessed on the evaluation sample. The best results are in bold. Calibration of the cutoffs (thresholds) was performed separately for each model based on the calibration set (i.e., validation set).

Table 8. Comparison of Predictive Power Across Subgroups of Restaurants

| Subgroup | Number of restaurants | | AUC |
	Open	Closed	
Overall	315	106	85.20
Age			
1st quartile	75	31	83.78
2nd quartile	66	39	84.60
3rd quartile	74	31	85.96
4th quartile	100	5	74.36
Chain status			
Affiliated	115	5	70.34
Independent	200	101	82.47
Restaurant category			
Fast food	121	35	88.36
Others	194	71	82.98

Note. The table shows the out-of-sample AUC values (i.e., based on the test set) for the VD-HMM in percentages.

compared various approaches, ranging from the simple mean (as used in practice) to machine learning methods to customized hidden Markov models. A classifier based on the mean rating attains an out-of-sample AUC of around 50% and thus does not exhibit predictive power. Machine learning methods based on feature engineering considerably improve the predictive power. However, the best performance is achieved by our customized HMM by modeling the complete rating sequence, with a gain in the out-of-sample AUC of 11.10 percentage points. Our novel VD-HMM achieves a balanced accuracy of 78.02%. In contrast, conventional classifiers from machine learning yield a balanced accuracy that is at least 5.97 percentage points lower and thus subpar. In simpler terms, our proposed model correctly identifies failures with a sensitivity of 67.86%; that is, 7 out of 10 failures are predicted correctly. Transitions to the at-risk state hint at service performance issues that, if they persist in the long run, are likely to lead to a business failure.

Based on our evidence, we find that this state is associated with a failure rate of 35.00% and, hence, should be seen as an early warning.

In order to enrich our quantitative results with qualitative insights, we compared the content of reviews across different states. We found that the content coincides with our identified latent states (see Table 11). Reviews from the well-running state reflect mostly positive experiences. The reviews from the at-risk state reveal that customer satisfaction is subject to considerable variance. Reviews from the bad-ratings-but-running state highlight consistent issues concerning customer satisfaction, and yet there are reasons why these restaurants remain in business (e.g., long opening hours, unique geographic location, drive-through service). These restaurants appear successful in drawing customers despite frequently poor ratings.

Our results reflect an exciting finding. One might have expected that the restaurants with the lowest ratings would have the highest failure risk. However, those restaurants often remain in business. There could be several explanations for this. According to the service-profit chain literature, quality is important, but so is resource input (e.g., Heskett et al. 1994, Kamakura et al. 2002, Mittal et al. 2005); that is, a low-rated restaurant could offer relatively low quality but at correspondingly low costs. Thus, it would not fail but remain in the market. For example, many fast-food chains (e.g., McDonalds, Burger King) employ this operating model of offering low-rated quality at low costs. This is also confirmed in our analysis, in which the bad-ratings-but-running state is common among restaurants from fast-food chains (see Table 11). Notwithstanding, such chains may have the advantage of running branches in exclusive locations, despite that these may have high costs (e.g., rent) and fail to turn a profit. Other branches of the chain could cross-subsidize those branches (Chevalier 2004). As a result, such a restaurant could also remain in the market despite poor ratings and high costs.

Table 9. Predicting Business Failures Early in Advance

Model	Features	1 month ahead	2 months ahead	3 months ahead	6 months ahead	9 months ahead	12 months ahead
Dynamic logit	Mean rating	51.44	51.56	50.97	48.69	48.44	53.84
Random forest	Mean rating	53.36	48.54	50.62	49.40	54.12	53.01
Dynamic logit	Rating dimensions	63.59	63.79	63.78	62.40	62.25	64.23
Random forest	Rating dimensions	54.23	53.05	55.72	56.67	54.45	57.66
Dynamic logit	Business covariates	74.00	73.84	74.31	74.79	74.98	75.45
Random forest	Business covariates	73.30	73.58	74.04	74.35	75.47	75.40
Dynamic logit	All	79.99	79.86	80.09	79.45	79.36	79.74
Random forest	All	79.74	81.00	81.87	80.87	80.12	80.96
Proposed VD-HMM		**85.05****	**84.95***	**84.43***	**83.67***	**82.90***	**81.73**

Notes. This table compares the out-of-sample performance to make early predictions of failures. The VD-HMM exhibits superior performance for each of the considered forecast horizons. It is further reported whether the improvement over the best machine learning baseline is statistically significant. The table shows out-of-sample AUC values (i.e., based on the test set) in percentages. The best model results are in bold.
  *$p < 0.05$; ** $p < 0.01$.

Table 10. Lead Time of the At-Risk State Until a Business Failure (Quantiles; in Weeks)

25%	50%	75%
31.96	78.43	136.65

6.2. Limitations and Potential for Future Research

As with any research, our results are subject to several limitations. Although our model allows us to separate different risk states and derive early warnings that alert service marketers *when* to intervene, our model does not tell service marketers *how* to reach a better state. We also refrain from claiming causal relationships between ratings and business failures, as our purpose is not to explain but to predict. Ratings might be subject to biases and, despite this, a favorable prediction performance achieved. Another possible limitation of our study is that we have focused on restaurants. Restaurants are certainly an important sector within the service industry and face considerable challenges (Luo and Stark 2015); however, future studies may look at other service sectors.

Our work opens several avenues for future research. One direction would be to improve the accuracy of our risk assessment. This could be achieved by considering further predictors. For instance, one may further extend our model to accommodate expert ratings (e.g., Michelin stars). In order to do this, one would simply need to add another emission component (see Ascarza and Hardie 2013), so that both customer and expert ratings are driven by the same latent dynamics. Although we already consider review sentiment, it might be interesting to combine our model with text mining frameworks (e.g., Zhong and Schweidel 2020). This would allow marketers to detect change points in review language, and thus offer interpretability through qualitative insights. A different research direction would involve diving deeper into the underlying mechanism of why restaurants fail.

Here, our model provides a starting point by showing that rating dynamics are an important determinant of business failures, whereas many other variables were not significantly linked with failure risk. Future research could also assess the effectiveness of different marketing interventions by modeling the treatment effect on the latent state behind service performance. This could allow marketers to quantitatively discern which interventions only improve customer satisfaction in terms of ratings and which are actually beneficial for survival (i.e., which interventions are effective in rectifying an at-risk state), thereby guiding how marketing efforts can be directed effectively.

6.3. Implications for Academia and Management

From a theoretical perspective, our research contributes to the existing literature on online ratings. Previous work has already shown a positive correlation between online ratings and business performance, often measured by sales (e.g., Godes and Mayzlin 2004, Chevalier and Mayzlin 2006, Liu 2006, Dellarocas et al. 2007, Chintagunta et al. 2010, Zhu and Zhang 2010, Archak et al. 2011; see Appendix A in the e-companion for a summary of relevant literature). We extend these valuable contributions by establishing a new relationship between customer satisfaction and business performance. Unlike earlier research, we do not measure sales but, rather, business failures. To the best of our knowledge, this paper is the first to evaluate the predictive power of online ratings as an early warning for business failures.

By suggesting a variable-duration HMM, we extend the growing literature on HMMs in the field of marketing. Previous literature has used HMMs to study different aspects of customer dynamics (e.g., Netzer et al. 2008, Montoya et al. 2010, Schweidel et al. 2011, Ascarza and Hardie 2013, Schwartz et al. 2014, Zhang et al. 2014b, Ascarza et al. 2018, Montoya and Gonzalez 2019). However, traditional HMMs are based on

Table 11. Sample Reviews

State	Rating	Review
(1) Well running	4	"Really nice staff and the place is clean."
(2) At risk	3	"My 'salad' had only 3 of the 7 veggies the menu promised, and a sorry portion at that. And while the balsamic vinaigrette was tasty, it didn't make up for my poor excuse for a salad."
(2) At risk	1	"I usually give second chances but this place was outright horrible! This place has bad, awkward, rude service & the food was between mediocre & bad."
(3) Bad ratings but running	1	"Kind of annoyed with them! I ordered an Oreo mcflurry with hot fudge Paid for my ice cream but no hot fudge on it Asked for and was told they were out."

the Markov property, according to which transitions can depend only on the previous state, whereas our VD-HMM considers the duration of latent states. This allows us to model effects where future dynamics are influenced by longer exposure to a latent state. This is likely to aid future research in marketing when similar dynamics must be modeled (e.g., to capture learning effects or retention).

Our model has also practical implications for managers, customers, and platform providers. First, *managers* can use our model to inform their decision making because the model provides an early warning system for business failures. They can use the model to predict business failures and, based on the estimated risk, plan timely interventions. Second, the results of the model could also benefit *customers*. Assuming a rating platform decides to display a restaurant's state on its site, customers could factor this into their decision. For example, they could decide to avoid bad-ratings-but-running restaurants or support a favored restaurant when it is in a risky state. Third, *platform providers* could also benefit from our model. They could choose to offer our model as another premium service. For example, Yelp for Business already offers several paid features for business intelligence.[10] This allows restaurant owners to gain data-driven insights into their business performance, thereby informing corporate strategy and operations (including decisions pertaining to credit as well as investors or lenders who might provide investment). A risk analysis could complement these services.

6.4. Conclusion

In this paper, we develop a hidden Markov model for predicting business failures from customer ratings. Using restaurant ratings, our model derives three latent states: well running, bad ratings but running, and at risk. Businesses in the at-risk state are associated with a higher risk of failure. Our model predicts business failures months before they occur, giving marketers sufficient time to plan interventions and, ultimately, prevent businesses from failing.

Endnotes

1 There are two common approaches with regard to how time is captured in HMMs (see Ascarza and Hardie 2013). One approach is to sum usage statistics per time interval. Another is to model the actual events (e.g., Montgomery et al. 2004, Ding et al. 2015). Therein, the time between events is inserted in the model to account for calendar time. It makes the assumption that latent states can only change at each rating event. The second approach is followed in this paper for practical reasons. It allows business owners to update the model and thus the prediction whenever new information becomes available, that is, whenever a new rating is submitted. In addition, it treats the time difference between two ratings as informative, it circumvents the need for feature engineering (e.g., as we shall see later, the mean rating does not predict business failure and is thus not effective for this purpose), and it allows us to incorporate all information accompanying a rating event (e.g., the review sentiment).

2 The prediction setting in Ding et al. (2015) and Montgomery et al. (2004) is the following: an HMM receives an observation sequence from an *unseen* entity as input and the model then outputs a classification (i.e., a single score) for the sequence, that is, for a new entity. This is different from path forecasting (e.g., Gopalakrishnan et al. 2018), where the model operates on the *same* entity at both training and deployment. In contrast to that, we are interested in a setting where a model generalizes *across* businesses, as this is crucial to provide value in management practice. This allows us to identify which (and when) businesses fail.

3 Using duration (in days) yields an equivalent approach. The reason is that we later also insert the time difference between ratings into the model (in addition to d_j) and, thus, can be transformed into one another. Nevertheless, we implemented our model where d_j was measured as the number of days. However, this led to overfitting and, thus, inferior results.

4 In theory, the failure of a business could have been modeled through an absorbing state. We experimented with this approach as part of our robustness checks, but it results in an inferior prediction performance. This can be explained by the fact that businesses can recover after reaching an at-risk state. Through improvements in service quality, businesses can eventually earn positive customer feedback again; thus, the at-risk state does not necessarily need to be terminal.

5 We also tested different structural assumptions in the transition matrix. However, this proved not to be beneficial. For example, Abhishek et al. (2012) encoded a funnel structure in which transitions could occur only between neighboring states. We experimented with their approach; however, it resulted in an inferior model fit.

6 The start of the study period was set to 2010, as several key platform features (e.g., check-ins and other community features) were introduced in 2010, and, hence, data were available only from 2010 onward. Robustness checks comparing the prediction performance across different time frames led to conclusive findings.

7 We performed a robustness check in which these events were labeled as failures, which resulted in an overall similar discriminatory power of our model.

8 Our prediction setting is different from path forecasting, which operates only *within* an entity (e.g., Gopalakrishnan et al. 2018). In contrast, we leverage the cross-sectional structure of our data in order to learn patterns of failures *across* restaurants. This allows us support managers with early warnings for their own businesses.

9 The cutoffs (thresholds) are used to translate the failure probability $\pi_i(j) \in [0,1]$ (see Equation (3)) into a binary prediction $\hat{F}_i \in \{0,1\}$ of the failure of restaurant i. Hence, the model specific cutoff is given as a real number $\delta \in [0,1]$ such that

$$\hat{F}_i = \begin{cases} 0, & \text{if } \pi_i(j) < \delta, \\ 1, & \text{if } \pi_i(j) \geq \delta. \end{cases} \tag{8}$$

10 See https://www.yelp.com/knowledge for details on Yelp's data-driven insights.

References

Abhishek V, Fader P, Hosanagar K (2012) Media exposure through the funnel: A model of multi-stage attribution. Preprint, submitted October 8, https://dx.doi.org/10.2139/ssrn.2158421.

Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Finance* 23(4):589–609.

Archak N, Ghose A, Ipeirotis PG (2011) Deriving the pricing power of product features by mining consumer reviews. *Management Sci.* 57(8):1485–1509.

Argenti J (1976) *Corporate Collapse: The Causes and Symptoms* (John Wiley and Sons, New York).

Ascarza E, Hardie BGS (2013) A joint model of usage and churn in contractual settings. *Marketing Sci.* 32(4):570–590.

Ascarza E, Netzer O, Hardie BGS (2018) Some customers would rather leave without saying goodbye. *Marketing Sci.* 37(1):54–77.

Asubonteng P, McCleary KJ, Swan JE (1996) SERVQUAL revisited: A critical review of service quality. *J. Services Marketing* 10(6):62–81.

Babić Rosario A, Sotgiu F, de Valck K, Bijmolt TH (2016) The effect of electronic word of mouth on sales: A meta-analytic review of platform, product, and metric factors. *J. Marketing Res.* 53(3):297–318.

Barbu VS, Limnios N (2008) *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications: Their Use in Reliability and DNA Analysis* (Springer, New York).

Barron K, Kung E, Proserpio D (2021) The effect of home-sharing on house prices and rents: Evidence from Airbnb. *Marketing Sci.* 40(1):23–47.

Bayou M, Bennet LB (1992) Profitability analysis for table-service restaurants. *Cornell Hotel Restaurant Admin. Quart.* 33(2):49–55.

Beaver WH (1966) Financial ratios as predictors of failure. *J. Accounting Res.* 4:71–111.

Berger J, Humphreys A, Ludwig S, Moe WW, Netzer O, Schweidel DA (2020) Uniting the tribes: Using text for marketing insight. *J. Marketing* 84(1):1–25.

Buttle F (1996) SERVQUAL: Review, critique, research agenda. *Eur. J. Marketing* 30(1):8–32.

Camillo AA, Connolly DJ, Kim WG (2008) Success and failure in Northern California. *Cornell Hospitality Quart.* 49(4):364–380.

Chen Z, Lurie NH (2013) Temporal contiguity and negativity bias in the impact of online word of mouth. *J. Marketing Res.* 50(4):463–476.

Chevalier JA (2004) What do we know about cross-subsidization? evidence from merging firms. *Adv. Econom. Anal. Policy* 4(1):1–27.

Chevalier JA, Mayzlin D (2006) The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* 43(3):345–354.

Chiappa S (2014) Explicit-duration Markov switching models. *Foundations Trends Machine Learn.* 7(6):803–886.

Chintagunta PK, Gopinath S, Venkataraman S (2010) The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* 29(5):944–957.

Culotta A, Cutler J (2016) Mining brand perceptions from Twitter social networks. *Marketing Sci.* 35(3):343–362.

Dellarocas C, Narayan R (2006) A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Statist. Sci.* 21(2):277–285.

Dellarocas C, Zhang XM, Awad NF (2007) Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *J. Interactive Marketing* 21(4):23–45.

Ding AW, Li S, Chatterjee P (2015) Learning user real-time intent for optimal dynamic web page transformation. *Inform. Systems Res.* 26(2):339–359.

Fader PS, Winer RS (2012) Introduction to the special issue on the emergence and impact of user-generated content. *Marketing Sci.* 31(3):369–371.

Fader PS, Hardie BG, Liu Y, Davin J, Steenburgh T (2018) "How to project customer retention" revisited: The role of duration dependence. *J. Interactive Marketing* 43:1–16.

Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. *Statist. Comput.* 24(6):997–1016.

Godes D, Mayzlin D (2004) Using online conversations to study word-of-mouth communication. *Marketing Sci.* 23(4):545–560.

Gopalakrishnan A, Bradlow E, Fader P (2018) Limited dynamic forecasting of hidden Markov models. Preprint, submitted July 13, https://dx.doi.org/10.2139/ssrn.3206425.

Hatt T, Feuerriegel S (2020) Early detection of user exits from clickstream data: A Markov modulated marked point process model. Huang Y, King I, Liu TY, van Steen M, eds. *Proc. Web Conference* (Association for Computing Machinery, New York), 1671–1681.

Helsen K, Schmittlein DC (1993) Analyzing duration times in marketing: Evidence for the effectiveness of hazard rate models. *Marketing Sci.* 12(4):395–414.

Henderson AD (1999) Firm strategy and age dependence: A contingent view of the liabilities of newness, adolescence, and obsolescence. *Admin. Sci. Quart.* 44(2):281–314.

Heskett JL, Jones TO, Loveman GW, Sasser WE, Schlesinger LA (1994) Putting the service-profit chain to work. *Harvard Bus. Rev.* 2:164–170.

Ho YCC, Wu J, Tan Y (2017) Disconfirmation effect on online rating behavior: A structural model. *Inform. Systems Res.* 28(3):626–642.

Ittner CD, Larcker DF (2003) Coming up short on nonfinancial performance measurement. *Harvrad Bus. Rev.* 81(11):88–95.

Kamakura WA, Mittal V, de Rosa F, Mazzon JA (2002) Assessing the service-profit chain. *Marketing Sci.* 21(3):294–317.

Ladhari R (2009) A review of twenty years of SERVQUAL research. *Internat. J. Quality Service Sci.* 1(2):172–198.

Laitinen EK (1991) Financial ratios and different failure processes. *J. Bus. Finance Accounting* 18(5):649–673.

Lee TY, Bradlow ET (2011) Automated marketing research using online customer reviews. *J. Marketing Res.* 48(5):881–894.

Lee YJ, Hosanagar K, Tan Y (2015) Do I follow my friends or the crowd? Information cascades in online movie ratings. *Management Sci.* 61(9):2241–2258.

Lervik-Olsen L, Witell L, Gustafsson A (2014) Turning customer satisfaction measurements into action. *J. Service Management* 25(4):556–571.

Li S, Sun B, Montgomery AL (2011) Cross-selling the right product to the right customer at the right time. *J. Marketing Res.* 48(4):683–700.

Liu Y (2006) Word of mouth for movies: Its dynamics and impact on box office revenue. *J. Marketing* 70(3):74–89.

Luca M (2011) Reviews, reputation, and revenue: The case of Yelp.com. Preprint, submitted September 16, http://dx.doi.org/10.2139/ssrn.1928601.

Luca M, Zervas G (2016) Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Sci.* 62(12):3412–3427.

Ludwig S, de Ruyter K, Friedman M, Brüggen EC, Wetzels M, Pfann G (2013) More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *J. Marketing* 77(1):87–103.

Luo T, Stark PB (2015) Nine out of 10 restaurants fail? Check, please. *Significance* 12(2):25–29.

Lussier RN (1995) A nonfinancial business success versus failure prediction model for young firms. *J. Small Bus. Management* 33(1):8–20.

Mahajan V, Srinivasan R, Wind J (2002) The dot.com retail failures of 2000: Were there any winners? *J. Acad. Marketing Sci.* 30(4):474–486.

Mittal V, Anderson EW, Sayrak A, Tadikamalla P (2005) Dual emphasis and the long-term financial impact of customer satisfaction. *Marketing Sci.* 24(4):544–555.

Moe WW, Schweidel DA (2012) Online product opinions: Incidence, evaluation, and evolution. *Marketing Sci.* 31(3):372–386.

Moe WW, Trusov M (2011) The value of social dynamics in online product ratings forums. *J. Marketing Res.* 48(3):444–456.

Montgomery AL, Li S, Srinivasan K, Liechty JC (2004) Modeling online browsing and path analysis using clickstream data. *Marketing Sci.* 23(4):579–595.

Montoya R, Gonzalez C (2019) A hidden Markov model to detect on-shelf out-of-stocks using point-of-sale data. *Manufacturing Serv. Oper. Management* 21(4):932–948.

Montoya R, Netzer O, Jedidi K (2010) Dynamic allocation of pharmaceutical detailing and sampling for long-term profitability. *Marketing Sci.* 29(5):909–924.

Moon S, Kamakura WA, Ledolter J (2007) Estimating promotion response when competitive promotions are unobservable. *J. Marketing Res.* 44(3):503–515.

Mun SG, Jang S (2018) Restaurant operating expenses and their effects on profitability enhancement. *Internat. J. Hospitality Management* 71:68–76.

Murphy KP (2002) *Hidden semi-Markov models (HSMMs).* Working paper, University of British Columbia, Canada.

Murphy KP (2012) *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA).

Netzer O, Lattin JM, Srinivasan V (2008) A hidden Markov model of customer relationship dynamics. *Marketing Sci.* 27(2):185–204.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Sci.* 31(3):521–543.

Ohlson JA (1980) Financial ratios and the probabilistic prediction of bankruptcy. *J. Accounting Res.* 18(1):109–131.

Olsen M, Bellas C, Kish LV (1983) Improving the prediction of restaurant failure through ratio analysis. *Internat. J. Hospitality Management* 2(4):187–193.

Parasuraman A, Zeithaml VA, Berry LL (1988) SERVQUAL: A multi-item scale for measuring consumer perceptions of service quality. *J. Retailing* 64(1):12–40.

Parsa HG, Gregory A, Terry M (2011a) Why do restaurants fail? Part III: An analysis of macro and micro factors. *Working paper, Dick Pope Sr. Institute for Tourism Studies,* University of Central Florida, Orlando.

Parsa HG, Self J, Sydnor-Busso S, Yoon HJ (2011b) Why restaurants fail? Part II—The impact of affiliation, location, and size on restaurant failures: Results from a survival analysis. *J. Foodservice Bus. Res.* 14(4):360–379.

Parsa HG, van der Rest JPI, Smith SR, Parsa RA, Bujisic M (2015) Why restaurants fail? Part IV: The relationship between restaurant failures and demographic factors. *Cornell Hospitality Quart.* 56(1):80–90.

Pretorius M (2009) Defining business decline, failure and turnaround: A content analysis. *Southern African J. Entrepreneurship Small Bus. Management.* 2(1):1–16.

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77(2):257–286.

Rust RT, Zahorik AJ, Keiningham TL (1995) Return on quality (ROQ): Making service quality financially accountable. *J. Marketing* 59(2):58–70.

Sarkar S, Sriram RS (2001) Bayesian models for early warning of bank failures. *Management Sci.* 47(11):1457–1475.

Schneider C, Weinmann M, Mohr PN, vom Brocke J (2021) When the stars shine too bright: The influence of multidimensional ratings on online consumer ratings. *Management Sci.* 67(6): 3871–3898.

Schwartz EM, Bradlow ET, Fader PS (2014) Model selection using database characteristics: Developing a classification tree for longitudinal incidence data. *Marketing Sci.* 33(2):188–205.

Schweidel DA, Bradlow ET, Fader PS (2011) Portfolio dynamics for customers of a multiservice provider. *Management Sci.* 57(3): 471–486.

Schweidel DA, Fader PS, Bradlow ET (2008) Understanding service retention within and across cohorts using limited information. *J. Marketing* 72(1):82–94.

Sharma S, Mahajan V (1980) Early warning indicators of business failure. *J. Marketing* 44(4):80–89.

Sismeiro C, Bucklin RE (2004) Modeling purchase behavior at an e-commerce web site: A task-completion approach. *J. Marketing Res.* 41(3):306–323.

Sonnier GP, McAlister L, Rutz OJ (2011) A dynamic model of the effect of online communications on firm sales. *Marketing Sci.* 30 (4):702–716.

Swaminathan A (1996) Environmental conditions at founding and organizational mortality: A trial-by-fire model. *Acad. Management J.* 39(5):1350–1377.

Tirunillai S, Tellis GJ (2012) Does chatter really matter? Dynamics of user-generated content and stock performance. *Marketing Sci.* 31(2):198–215.

Toubia O, Stephen AT (2013) Intrinsic vs. image-related utility in social media: Why do people contribute content to Twitter? *Marketing Sci.* 32(3):368–392.

Wang L, Gopal R, Shankar R, Pancras J (2015) On the brink: Predicting business failure with mobile location-based checkins. *Decision Support Systems* 76:3–13.

Yoon E, Jang S (2005) The effect of financial leverage on profitability and risk of restaurant firms. *J. Hospitality Financial Management* 13(1):35–47.

Yu SZ (2016) *Hidden Semi-Markov Models: Theory, Algorithms and Applications* (Elsevier, Amsterdam, Netherlands).

Zhang JZ, Netzer O, Ansari A (2014a) Dynamic targeted pricing in B2B relationships. *Marketing Sci.* 33(3):317–337.

Zhang KZ, Zhao SJ, Cheung CM, Lee MK (2014b) Examining the influence of online reviews on consumers' decision-making: A heuristic–systematic model. *Decision Support Systems* 67 (November):78–89.

Zhong N, Schweidel DA (2020) Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Sci.* 39(4):827–846.

Zhu F, Zhang XM (2010) Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *J. Marketing* 74(2):133–148.