# Model Selection Using Database Characteristics:

# Developing a Classification Tree for Longitudinal Incidence Data

Eric M. Schwartz

Eric T. Bradlow

Peter S. Fader *

* Eric M. Schwartz is an Assistant Professor of Marketing at the Stephen M. Ross School of Business at the University of Michigan. Eric T. Bradlow is the K. P. Chao Professor, Professor of Marketing, Statistics and Education, Vice-Dean and Director of Wharton Doctoral Programs, and Co-Director of the Wharton Customer Analytics Initiative, at the University of Pennsylvania. Peter S. Fader is the Frances and Pei-Yua Chia Professor, Professor of Marketing, and Co-Director of the Wharton Customer Analytics Initiative, at the University of Pennsylvania. The first author thanks the Wharton Risk Management and Decision Processes Center for its support through the Russell Ackoff Doctoral Student Fellowships, as well as Eva Ascarza, Michael Braun, Oded Netzer, and Kenneth Shirley for their useful comments. The authors all thank Yao Zhang, who contributed to a related working paper "Children of the HMM: Tracking Longitudinal Behavior at Hulu.com" which was presented at the 2010 Marketing Science Conference. All correspondence on this manuscript should be addressed to Eric M. Schwartz, ericmsch@umich.edu, 734-936-5042; 701 Tappan Ave., R5468, Ann Arbor, MI 48109-1234.

**Model Selection Using Database Characteristics:**
**Developing a Classification Tree for Longitudinal Incidence Data**


## Abstract

When managers and researchers encounter a dataset, they typically ask two key questions: (1) which model (from a candidate set) should I use? and (2) if I use a particular model, when is it going to likely work well for my business goal? This research addresses those two questions, and provides a rule, i.e., a decision tree, for data analysts to portend the "winning model" *before* having to fit any of them for longitudinal incidence data. We characterize datasets based on managerially relevant (and easy-to-compute) summary statistics, and we use classification techniques from machine learning to provide a decision tree that recommends when to use which model. By doing the "legwork" of obtaining this decision tree for model selection, we provide a time-saving tool to analysts. We illustrate this method for a common marketing problem (i.e., forecasting repeat purchasing incidence for a cohort of new customers) and demonstrate the method's ability to discriminate among an integrated family of a hidden Markov model (HMM) and its constrained variants. We observe a strong ability for dataset characteristics to guide the choice of the most appropriate model, and we observe that some model features (e.g., the "back-and-forth" migration between latent states) are more important to accommodate than others (e.g., the inclusion of an "off" state with no activity). We also demonstrate the method's broad potential by providing a general "recipe" for researchers to replicate this kind of model classification task in other managerial contexts (outside of repeat purchasing incidence data and the HMM framework).



**Keywords:** model selection, machine learning, data science, business intelligence, hidden Markov models, classification tree, random forest, posterior predictive model checking, hierarchical Bayesian methods, forecasting.

## 1 Introduction

The explosion in technology-enabled data collection has changed the focus of marketing modelers away from aggregated data at the store- or market-level towards more granular, panel-oriented, data structures and associated statistical methodologies. Companies have reduced their reliance on "rolled up" data provided by syndicated vendors (e.g., IRI, Nielsen) and now build more of their analytics around customer-level longitudinal patterns that they can obtain from their own internal operations. But while this increased reliance on "site-centric" data (Zheng et al. 2011) offers a number of meaningful benefits to the firm, it also comes with some potential costs to the researcher.

First, site-centric data provides a detailed description of each customer's stream of purchases (and other actions that the firm can measure directly), but it often lacks information about marketing variables, competitive tactics, and other potential "drivers" of the behavior(s) of interest (Donkers et al. 2007; Schweidel et al. 2008) that are typically provided by a third-party firm and often difficult to link to purchase data. So many firms are focusing their decision-making efforts around the flow of incidence activities, i.e., the timing and nature of each transaction, which is very rich, but also quite different than the inputs used in more traditional marketing-mix models (Hanssens et al. 2005).

Second, this detailed stream of incidence actions can be characterized by an ever-larger swath of mathematical models. That is, increased granularity comes with the potential for increased model complexity and hence a more difficult model selection problem than faced by previous generations of researchers, who often relied on relatively standard model specifications (Cooper and Nakanishi 1988; Wittink et al. 1988) that were sufficient for the relatively standard data structures that were made available by a small set of third-party data providers.

With this "data evolution" in mind, consider a business intelligence manager for an e-commerce firm who is examining panel data from three recent product launches (Figure 1). Her goal is to project repeat purchase patterns for each dataset, as her company's production, marketing, and customer relationship management (CRM) activities depend on an accurate forecast. How should she choose which statistical model is most appropriate for each product's dataset? She could run a number of different panel-oriented incidence models and choose the

one that fits each dataset best; but, a series of separate "model bake-offs" would be a highly inefficient process and would offer no assurances that the chosen model(s) will be best-suited for forecasting purposes of similar datasets. Instead, are there clues in each dataset that might help her make the right choices without having to run an array of models over and over again for each new dataset? Can we look at many datasets and model performance to extract general rules about when to use which model? That is the goal of this paper: we want to help managers choose among competing longitudinal incidence models, based only on observed dataset-level summary statistics, i.e., *database characteristics*, before they need to run any models.

[INSERT FIGURE 1 ABOUT HERE]

We will create a "decision tree" that can guide the manager towards the most appropriate model specification for a given dataset, based only on observable (and easy to compute) summary statistics on that dataset. In other words, we will do the "upfront work" so that the decision tree is a time-saving tool for other analysts. We recognize that each dataset consists of a mix of heterogeneous customers who may go through different kinds of dynamic purchasing patterns over time, and we want to identify the most suitable model specification to capture these within- and across-customer sources of variation. However, we do not use the database characteristics directly in our models to predict future purchasing (i.e. we don't treat them as X variables in a statistical model), but instead we use them to help identify the best model (chosen from a class of different latent-state model specifications), which can be used for forecasting and other diagnostic purposes.

For instance, referring back to Figure 1, Dataset A's steadily declining sales may indicate that latent customer attrition is prevalent but occurs at different rates for different customers, so a "buy till you die" model such as the Pareto/NBD (Schmittlein et al. 1987) or the BG/BB (beta-geometric beta-binomial; Fader et al. 2010) might be appropriate. In contrast, the sales for Dataset B seem to show a substantial rise towards the end of the observation period, so a hidden Markov model (HMM), in which customers move back and forth between different states of purchasing propensities (Liechty et al. 2003; Netzer et al. 2008), might be the best model to employ for forecasting purposes. Finally, the sales curve for Dataset C is harder to classify as a "buy till you die" or an HMM-type pattern - it seems to reflect elements of both

specifications. Perhaps we need a hybrid version of these two models to capture and project it.

While there are innumerable models that could be viable candidates for this kind of longitudinal incidence data, we choose a particular set that are tightly connected to each other but still very flexible. The models we consider are the HMM and three different constrained variants of it (including the BG/BB). Since they are part of an integrated family, they offer an opportunity to detect when each underlying model component (in this case the presence of an absorbing state and/or the need for a "no-purchase" state) is worth including or "turning off." This provides added insight to the analyst about the nature of their customer dynamics, above and beyond simplified model implementation and improved model performance.

For this context (i.e., repeat-transaction incidence data; HMM and its constrained variants), we do all of the "leg work" for the analyst. We run an array of constrained and unconstrained HMM models on dozens of synthetic datasets, generated to broadly represent the kinds of patterns that are likely to occur in real-world settings. While this is computationally expensive initially (a high "upfront cost" for us as the researcher), it yields significant savings for the downstream user - the manager simply follows our advice and selects the most appropriate model given the nature of her dataset, and runs it, "the winner" and not the entire class of models.

The focal managerial criterion we use to select among models is the forecast error for each cohort's purchases; so the winning model has the minimum mean absolute error in a holdout period. We use well-established machine-learning methods known as *classification and regression trees* (CART) and *random forests*, to derive general rules to suggest which model to use under different circumstances, based entirely on observed (and managerially meaningful) patterns in the customer-base data. The database characteristics that turn out to be most important (in our setting) include the nature of the decline in cohort-level sales over time as well as purchase concentration (e.g., the "80:20 rule") across customers.

Since the development of the decision tree is our key contribution, the structure of the paper centers around it. There are three "ingredients" for the classification approach, and we devote a section of the paper to each one: the candidate models in Section 2, the database characteristics in Section 3, and the performance criterion to determine the "winning" model

for each dataset in Section 4. Putting these three ingredients together, we create the decision tree in Section 5 and focus on its interpretation, validation, and managerial implications.

Although we perform our analysis for a specific data/modeling context (albeit an important one in today's marketing environment), the same basic "recipe" developed here can be applied to many other settings. Thus, we formalize our approach as a more general methodology in the Appendix, using the same ingredients outlined above: a set of models, database characteristics, and a selection criterion (i.e., performance or error measure with loss function). We now begin the process of laying out these elements to build our decision tree.

## 2 Which models to consider? The HMM and its constrained variants

The decision tree recommends which model to use for a given dataset, but we have to provide a consideration set of models: the HMM and its constrained variants. Why do we consider this class? First, they are appropriate for this popular context of understanding and projecting repeat-purchase patterns of a cohort of customers using longitudinal incidence data (Liechty et al. 2003; Montgomery et al. 2004; Montoya et al. 2010; Netzer et al. 2008; Schweidel et al. 2011). Second, these models cover a wide range of underlying "stories" of customer behavior leading to different observable data patterns. This helps us achieve the goal of the paper: establish the link between dataset-level summaries and model performance.

Third, these models form an integrated family; that is, each model is a constrained or unconstrained version of another in the set. Some of these are established yet seemingly unrelated models, such as "buy till you die" and latent-class models, among others. But they are all special cases of the HMM. These connections have only been partially explored and in an ad-hoc manner in the previous literature, as we discuss below (Netzer et al. 2008; Schweidel et al. 2011). However, the extra insight that we provide is that the four variants of the HMM that we consider are described by two model components each with two levels (as seen in the $2 \times 2$ discussed in Figure 2). So the decision tree not only recommends a specific model, it also emphasizes the presence/absence of more general model components, thereby adding more insight and comparability across datasets.

The unconstrained HMM used here has two states, and the within-state purchase likeli-

hoods are repeated Bernoulli trials by individuals who can begin the calibration period in State 1 or State 2. We allow for unobserved continuous heterogeneity for both the within-state purchase propensities as well as the between-state transition probabilities. Formally, we let $y_{it} = 1$ for day $t$ if the customer $i$ purchased and $y_{it} = 0$ otherwise.[1] Then,

$$y_{it} \sim \begin{cases} \text{Bernoulli}(p_{1i}) \text{ if in State } 1 \text{ , } Z_{it} = 1 \\ \text{Bernoulli}(p_{2i}) \text{ if in State } 2 \text{ , } Z_{it} = 2 \text{ ,} \end{cases} \tag{1}$$

where the latent-state variable, $Z_{it}$, indicates which state a customer occupies on each day. The individual-level parameters of the HMM are the within-state propensities, $\mathbf{p}_i$, and the transition probability matrix, $\mathbf{\Theta}_i$. That is,

$$\mathbf{p}_i = (p_{1i}, p_{2i}) \text{ and } \mathbf{\Theta}_i = \begin{pmatrix} 1 - \theta_{12i} & \theta_{12i} \\ \theta_{21i} & 1 - \theta_{21i} \end{pmatrix}. \tag{2}$$

We let the initial state membership be a population-level parameter, and any individual can start in State 1 with probability $\pi_1$ or State 2 with probability $1-\pi_1$. We assume independent beta distributions to allow for heterogeneity across individuals for the components of $\mathbf{p}_i$ and $\mathbf{\Theta}_i$.[2] Specifically, the prior distributions used are

$$\begin{aligned} p_{1i} &\sim \text{beta}(\mu_{p_1}, \phi_{p_1}), & p_{2i} &\sim \text{beta}(\mu_{p_2}, \phi_{p_2}), \\ \theta_{12i} &\sim \text{beta}(\mu_{\theta_{12}}, \phi_{\theta_{12}}), & \theta_{21i} &\sim \text{beta}(\mu_{\theta_{21}}, \phi_{\theta_{21}}), \end{aligned} \tag{3}$$

where $\mu = a/(a+b)$ is the mean and $\phi = 1/(a+b+1)$ is the polarization index of the beta distribution with shape parameters $a$ and $b$ (Sabavala and Morrison 1977). In general, for $S \geq 2$ states, each row $r$ of the transition probability matrix is a vector $(\theta_{r1i}, \ldots, \theta_{rSi}) \sim \text{Dirichlet}(\alpha_{r1}, \ldots, \alpha_{rS})$.[3] We distinguish between states by referring to State 1 as having a

[1] Without loss of generality, we use "day" to refer to the unit of discrete time, and "purchase" as the observed behavior of interest. It could be instead, for example, viewing online videos or not in a given week, donating or not in a given quarter, etc.

[2] The initial state-membership probability is assumed to be a population-level parameter due to the definition of a cohort of customers acquired at the same time. Additionally, the results are robust to using a logit-normal for heterogeneity on all individual parameters, and for allowing correlations among them.

[3] We use highly uninformative hyperpriors on the population-level parameters of the beta (or Dirichlet) distri-

within-state propensity at least as large as that of State 2, i.e., for each individual (i.e., $p_{1i} \geq p_{2i}$ for all $i$). This prevents the label-switching problem known to exist with latent-state models (Stephens 2000).

We highlight the off-diagonal entries of the transition probability matrix, since $\theta_{12i}$ denotes the probability an individual moves "forward" (State 1 to 2) and $\theta_{21i}$ represents the probability an individual moves "backward" (State 2 to 1). So not allowing backwards transitions is equivalent to making State 2 absorbing.

[INSERT FIGURE 2 ABOUT HERE]

Given this formulation of the unconstrained HMM, the three nested models emerge as we constrain either or both model components. To start, when we apply both constraints to all individuals, so there is an "off"' state ($p_{2i} = 0$) and "backwards" transitions are prohibited ($\theta_{21i} = 0$), the "buy till you die" BG/BB model emerges. [4]

Then, as we think about how the BG/BB model and HMM differ along these two dimensions, we can consider each of those dimensions separately (i.e., either $p_{2i} = 0$ or $\theta_{21i} = 0$). These constraints determine the two dimensions of Figure 2. When applying each of the two constraints separately, different models emerge (the off-diagonal cells of Figure 2), and each tells a distinct story of customer behavior.

When only $p_{2i} = 0$, the *On And Off* model (OF) emerges. Consumers can make back-and-forth transitions between an "on" state of activity and an "off" state of inactivity. Like the HMM, customers can make backwards transitions, but like the BG/BB, when in the "off" state, customers have no chance of activity. This kind of model has been explored in papers on Markov-modulated Poisson processes (Ma and Buschken 2011).

Alternatively, when only $\theta_{21i} = 0$, we get the *Hot Then Cold* model (HC). At any time customers can be either in a "hot" state (higher propensity to purchase) or a "cold" state (purchasing is less likely but still possible). Like the BG/BB, once the customer reaches the "cold" state she remains there (no backwards transitions), and like the HMM, in the "cold"

butions. For more details about the distributions used in the sampling procedure, see the Appendix.

[4] Unlike the BG/BB as in Fader et al. (2010), which assumes that all individuals start in the "alive" state, in our BG/BB specification we allow individuals to start in either state, according to initial state probability vector, $\pi$. The model utilized here is more general.

state purchasing is possible. The hot-then-cold ordering is informed by the prevalence of customer attrition, or at least, slowing down of transactions (in aggregate) that is common in most cohort-level datasets.[5] Such behavior appears in queuing theory models, such as phase-type distributions (Bladt and Neuts 2003; O'Cinneide 1990), and in marketing models (Fader et al. 2004; Schweidel and Fader 2009).

Past literature has noted how the latent-class model (Kamakura and Russell 1989) is a special case of an HMM ($\theta_{12i} = \theta_{21i} = 0$), and other work often utilizes a nested model with a "death" state (Netzer et al. 2008; Schweidel et al. 2011). However, the other links among the HMM and its constrained models (e.g., BG/BB, HC, and OF) that we consider have not been documented in full detail as an integrated framework with the $2 \times 2$ structure as described here.

Viewing the HMM and its constrained variants as an integrated family provides an opportunity to detect when (i.e., for which types of datasets) each model component is worth including. One may initially (but erroneously) think that the nested structure would guarantee that the more flexible HMM would perform at least as well as any of its constrained versions (with one or both model components shut off) on all model-performance criteria. But this is not guaranteed in practice. We illustrate that when forecasting repeat transactions out-of-sample, the more general model does not always beat its nested versions, and hence there is value in the decision tree provided in this paper.

The decision tree answers our key question is: for what kinds of database characteristics does each model perform best? To perform this classification, we need a range of different datasets generated from the $2 \times 2$ framework. We generate 64 synthetic datasets, each with $T = 30$ weeks of data in calibration (and 30 for holdout) and $N = 500$ customers, with considerable variation by simulating them from unconstrained and constrained versions of the HMM (i.e., to capture each of the sub-models as well as the full unconstrained HMM) with a

[5]For this reason, we do not consider a separate "Cold Then Hot" model, although the general HMM and OF specifications allow individual-level purchasing to "speed up" over time.

generous range of population-level parameters:

$$\begin{aligned} \mu_{p1} &\in [0.05, 0.50] & \mu_{p2} &\in [0.00, 0.10] \\ \mu_{\theta 12} &\in [0.10, 0.35] & \mu_{\theta 21} &\in [0.00, 0.25] \\ \phi_{p1}, \phi_{p2}, \phi_{\theta 1}, \phi_{\theta 2} &\in [0.10, 0.45] & \pi_1 &\in \{0.50, 1.00\}. \end{aligned} \tag{4}$$

We discuss these synthetic datasets and the variability across them in the next section. However, after creating these datasets, we put aside the data-generating process and describe them entirely by easy-to-compute and managerially relevant database characteristics, which we now cover in detail.

## 3 Selecting database characteristics

The decision tree is a tool that predicts which specification is likely to be the "winning" model by only looking at summary statistics of a particular database. So, just as we need a set of reasonable models to choose from, we also need a set of database characteristics to drive the choice process. But which database characteristics should we consider? We illustrate our process of identifying relevant database characteristics by returning to one of the opening examples, repeat purchasing for Dataset A. Before running any models, analysts frequently examine two typical displays of a cohort's purchasing behavior: a cross-sectional histogram of customer-level transactions and a longitudinal tracking plot of cohort-level purchases over time. These two graphs appear in Figure 3.

[INSERT FIGURE 3 ABOUT HERE]

What are the key features of each graph? We want to choose summaries that are both managerially relevant and easy to compute directly from these aggregate plots. We identify four summaries that offer a fairly complete characterization of each plot. For the histogram, we propose summaries to capture the nature of the "head" and the "tail" of the distribution, as well as its central tendency. For the tracking plot, we focus on the early and late trends in purchasing as the cohort ages and the trend's overall "bowed" shape, as well as the overall variability over time.

More specifically, Table 1 contains a listing of these measures, which we will use in our subsequent empirical analysis. While there is not an exact science to selecting these measures, we choose them here to represent central tendency (e.g., average frequency), higher moments (e.g., top percentile, purchase concentration 80:20-type rule), and trend behavior (e.g., steepness, shape, trend variability). We do not claim this list to be comprehensive, but these values vary widely and in systematic ways across the datasets generated by the HMM and its constrained versions.

The variation in these measures across databases is essential: it allows us to explicitly show the range of empirical patterns we consider here, and is required to obtain a meaningful classification tree linking these summaries to the model selection process. We illustrate some of this variation in values of these summary statistics for Datasets A, B, and C (Table 2).

[INSERT TABLE 2 ABOUT HERE]

It is interesting to see how the datasets are indistinguishable on some dimensions (e.g., frequency), quite distinct from each another on others (e.g., penetration), and occasionally exhibit pairwise similarities (e.g., late trend for Datasets B and C).

In most empirical settings, we think about the amount of information as being related to the number of observations within a dataset. But in this setting each dataset is reduced to a single observation described along multiple dimensions, i.e., the database characteristics described above. Thus, we construct a "dataset of datasets," a collection of 64 simulated datasets reflecting variation along the summary statistics and representing real-world datasets (Fader et al. 2010; Netzer et al. 2008; Schweidel et al. 2011). Specifically, we generate datasets from all possible combinations of the parameter values noted in Section 2 which allows us to reflect both the structural variation as well as the "natural randomness" that arises from simulating the purchases. Once the datasets are created, the true values of the population-level parameters are no longer taken into consideration.

Figure 4 shows the large variability along the values of the database characteristics across the simulated datasets. For instance, nearly half of the datasets have penetration rates between 40% and 70%. About 40% of them have a very steep declining trend (steeper than a drop in transactions equivalent to 15% of the cohort size), while others show some growth in

purchases for the cohort over time. Thus, we believe that, by selecting and creating datasets in this way, we will have avoided biasing our classification results to favor any particular model specification.

[INSERT FIGURE 4 ABOUT HERE]

To ensure that our chosen characteristics are explaining most of the meaningful variation across the collection datasets, we ran a principal components analysis and an exploratory factor analysis on an even larger set of summary statistics beyond the ones described earlier. We do not present the detailed results but note a few highlights. The principal components analysis indicates that 99% of the measured variation across the 64 datasets can be captured by six independent components. The loadings of the principal components analysis and the loadings of the exploratory factor analysis (with 5, 6, and 7 factors) all point to a very similar set of summary statistics, such as central tendency, concentration, and variation over time.

We also recognize that a number of these database characteristics are naturally correlated with each other. Some measures are quite independent (e.g., late trend and penetration, $r = 0.01$), but other pairs have correlations that are large and significant (e.g., average frequency and penetration, $r = 0.86$). While this kind of multicollinearity could be a serious problem in a typical regression-like model, it does not affect the classification tree and random forest approach since they are non-parametric methods designed specifically for (sequential) variable selection (Breiman et al. 1984; Breiman 2001a).

## 4 Assessing model performance

The final ingredient that goes into the classification tree is a rule for declaring a winning model for a given dataset. Here, we select a "winner" based on each model's ability to predict an important managerial quantity that is widely used for purchasing data due to its link to customer lifetime value and other profit measures: aggregate incremental sales over a holdout period. Specifically, we select an error measure that summarizes the time series of discrepancies between the model and the observed sales for each "MCMC world." We will look at the variability of the errors across "worlds" and also average the errors across the "worlds" to obtain a measure of the model's error for that dataset that integrates over the posterior uncertainty.

The error measure we use, mean absolute error (MAE), assumes a linear loss function and is frequently used for time-series data. In the more general formulation of this procedure (see Appendix), one can select any managerial quantity (replacing out-of-sample aggregate sales over time) and error measure with a different loss function (to replace MAE). While we present results using MAE for our context, our classification tree results are robust to alternative common summary error measures (e.g., mean absolute percent error, MAPE; root mean squared error, RMSE).[6]

Formally, we quantify performance as the degree to which the model-based posterior predictive distribution of out-of-sample aggregate sales is outlying with respect to the quantity's observed value. We assume the posterior distribution has been obtained using standard MCMC procedures (detailed in the Appendix), yielding posterior draws $g = 1, \ldots, G$. For dataset $k$, $y_{kt}^{\text{obs}}$ is number of the observed incremental transactions at time period $t$ and $y_{kmt}^{*(g)}$ is one replicate from model $m$'s corresponding posterior predictive distribution for that quantity (i.e., incremental transactions). Then for each posterior replicate $g$, we compute the mean absolute error,

$$d_{km}^{(g)} = \frac{1}{T} \sum_{t=1}^{T} \left| y_{kmt}^{*(g)} - y_{kt}^{\text{obs}} \right|. \tag{5}$$

We will use the values of $d_{km}^{(g)}$ in two ways. On the one hand, we will examine the average posterior MAE for all four models on each dataset to determine a single winner per dataset. On the other hand, in order to provide a more nuanced set of findings, we characterize the full posterior uncertainty of the MAE by computing the probability that each model has lowest value (i.e., proportion of times each model is the winner across the $G$ posterior replicates) as we would not want to overly penalize a model which is a "close second," for instance. We also use the latter directly in our classification tree.

[6]We note that our choice of error measure for model selection is in contrast to commonly used likelihood-based summary criteria, such as BIC and DIC (Montgomery et al. 2004; Montoya et al. 2010; Netzer et al. 2008; Schweidel et al. 2011). We use an empirical quantity for model selection since many scholars caution against using purely likelihood-based measures (Gelman and Rubin 1995), especially for latent-state models, such as the HMM and its variants, because one must face issues with unstable estimators, computation of the posterior distribution, and correction factors of the log-marginal likelihood (Chib 1995; Lenk 2009; Newton and Raftery 1994; Spiegelhalter et al. 2002).

Now, armed with a set of models (the HMM and its constrained variants), the in-sample database characteristics for each dataset (Table 1), and an error measure (out-of-sample MAE), we have all of the ingredients for the decision tree, which is described next.

## 5 When to use which model? A classification tree

We classify datasets to reveal how we can select the model with the best out-of-sample error by only using in-sample database characteristics. This enables us to answer the paper's central question: given a dataset's summary statistics, which model best fits the data?

The winning model, $m_k^{\text{Winner}(g)}$, for dataset $k$ and posterior world $g$ is determined by the identifying the model with the minimum error $d_{km}^{(g)}$ among all $M$ models,

$$m_k^{\text{Winner}(g)} = \text{argmin}_{m=1,\ldots,M} \left\{ d_{km}^{(g)}, \ldots, d_{kM}^{(g)} \right\}, \tag{6}$$

We use a classification tree to relate the identity of the winning model, $m_k^{\text{Winner}(g)}$, to the vector of database characteristics, $\tau(\mathbf{Y}_k^{\text{obs}})$. Given the performance of all $M$ models across all $K$ datasets and $G$ posterior replicates, we explain variations in the model performance (i.e., which model wins) as a function of the observed summaries of that dataset. Formally, we capture this relationship as follows,

$$\hat{m}_k^{\text{Winner}(g)} = \text{Tree} \left[ \tau(\mathbf{Y}_k^{\text{obs}}) \right], \tag{7}$$

where the function "Tree" denotes the classification tree predicting the winning model $\hat{m}_k^{\text{Winner}(g)}$ for each of the datasets $k = 1, \ldots, K$ and posterior world $g = 1, \ldots, G$.

The classification tree provides cutoff values of the dataset-level summary statistics to place entire datasets into "buckets." This classifies datasets in an easy-to-interpret manner. Each bucket of datasets has a similar profile of dataset-level summary statistics and similar patterns of model performance. Therefore, when a new dataset is encountered, it can be classified using this decision rule to identify which of the models will likely be most suited for it. This allows us to uncover relationships between observed patterns in the data and model fit that are easy to interpret while avoiding the need to make any additional assumptions about functional

form or error distributions common to ordinary regression models.

Additionally, our classification tree approach goes one step further because it also reflects the natural parameter and model uncertainty. We reflect that uncertainty since our Bayesian modeling approach provides the full posterior distribution of performance for each dataset-model pair. As a result, each case to be classified is unique to a particular posterior draw from a model run on a dataset. This means that the data to be used to construct the classification tree contains $G = 100$ model-based replicates of the $K = 64$ observed datasets. By using $G$ replicates of each set of observed dataset summaries (independent variables), we allow for $G$ different values of errors from each model-dataset pair, and hence each dataset has a distribution of different winning models (dependent variables), and therefore receives an appropriate number of "votes."

### 5.1 Classification tree

The classification tree in Figure 5 can be easily read by starting at the top and following a series of "if/then" decisions down to a terminal node at the bottom of each branch. These terminal nodes represent a group of datasets with the same observed summary statistic branch values (predictor variables). Each node has a recommended winning model but also displays the within-node winning percentages for each model (based on the number of posterior worlds in which each model had the lowest forecast error). Note that the "N" values in the tree sum up to 6,400 cases, reflecting the use of 100 posterior replicates for each of the 64 simulated datasets.

[INSERT FIGURE 5 ABOUT HERE]

Four database characteristics were selected by the classification tree's sequential variable selection algorithm as being diagnostic: early trend, late trend, concentration, and trend Gini (trend shape).

The early and late trend statistics reflect the change in transactions over each half of the calibration period (15 days in each half) expressed as a percentage of the total customer base (500 customers). The classification tree partitions early trend into three levels: *very steep* (steeper than a drop in daily transactions equivalent to 11% of the customer base), *moderately*

*steep* (a drop between 1% and 11%), and *relatively flat or positive* (a slope that is more positive than -1%). Late trend is partitioned into two levels, which we label *moderately steep* (a drop steeper than 3% of the total customer base) and *relatively flat or positive* (a slope that is more positive than -3%). Next, the split for concentration has a remarkable resemblance to the 80/20 rule. A dataset is either *highly concentrated* (more than 82% of purchases are made by the top 20% of customers) or *not highly concentrated*.

The trend Gini summary statistic reflects the shape of the curve. How much does the actual curve deviate from a line connecting the first and last days of the calibration period (i.e., how much area is there between curve and the trend line)? In other words, this measures the degree to which the curve is "bowed." The variable is split into a less bowed shape (close to linear with value less than 5%) and a more bowed shape (value greater than 5%). Negative values indicate that there is more area between the curve and the trend line that sits above the trend line than below the trend line.

We illustrate the use of the tree by returning to our three introductory datasets. Recall their database characteristics were shown in Table 2. We can trace how the tree classifies these datasets to illustrate exactly how a manager can use our decision tree. For instance, Dataset A exhibits a sales pattern that is downward sloping early on (steeper than -1%), not strongly downward sloping later on (equal to -3%), and over 82% of the purchases are made by the top 20% of customers. Thus, the classification tree recommends that it would be best modeled using the BG/BB, since that model provides the best out-of-sample forecast for 40% of the posterior replicates associated with the 15 different datasets that have similar values of database characteristics. (And indeed, the BG/BB does provide the best forecast for Dataset A, as we show below in Tables 3, 4, and 5.)

It is interesting to note the internal consistency of the tree. In particular, the precise value of the late trend for Dataset A is exactly the classification tree's cutoff value (-3%). So even if the dataset's late trend were just slightly less than that cutoff, the dataset would still fall into a node dominated by the BG/BB model (i.e., in the leftmost terminal node of the tree, the BG/BB is the best-performing model in 46% of posterior replicates).

For datasets with a declining early trend and a flat or increasing late trend, but without a high purchase concentration (fewer than 82% of the purchases made by the top 20% of customers), a different pattern emerges. Datasets B and C are two such examples, so they fall into two terminal nodes in this part of the tree. Datasets represented in this part of the tree show a strong need to allow for back-and-forth transition (HMM and OF). But within the back-and-forth pair, there is less certainty about which one wins.

Further splitting the datasets by trend Gini (trend shape) over the calibration period and by early trend one more time allows the analyst to better discriminate when each model is likely to perform better. For datasets with a more bowed shape (trend Gini greater than or equal to 5%) and a very steep early trend (steeper than 9%), such as Dataset B, the HMM wins with 45% of votes versus the OF with 32%. However, for others with a moderately steep early trend (between a 1% and 9% drop) and a more bowed shape, such as Dataset C, the OF wins with 55% of votes versus the HMM with 32%. Datasets B and C are therefore best classified by the HMM and OF, respectively.

The split on trend shape (trend Gini) and an additional split on early trend should be intuitive as the HMM is a more general model than the OF. As a result, the HMM can generate a wider range of patterns across datasets than the OF, due to the extra model flexibility (e.g., State 2 purchase probability is not necessarily zero). To understand this, keep in mind the patterns common to the datasets in this part of tree: not highly concentrated purchasing and flat or increasing late trend. On one hand, for less bowed-shaped curves, the OF has difficulty capturing a nearly linear pattern since the "off" state induces a moderate steep early drop. On the other hand, for markedly bow-shaped curves, the OF also has difficulty capturing both the very steep early declining trends and flat or increasing later trend. Capturing such an interaction among database characteristics is an advantage that CART methods has over traditional linear regression approaches.

### 5.2 Uncertainty in model performance

As we take a deeper dive into particular branches of the decision tree, we examine the uncertainty in model performance. That is, while the model with the lowest error is declared

the "winner," we describe how we one can utilize "votes" for each winning model by utilizing the full posterior from the Bayesian model output.

As an illustration, we return to Datasets A, B, and C to look at the comparative performance of the HMM and its variants from the $2 \times 2$ framework. The average performance seen in the plots of Figure 6 are quantified in Table 3.

[INSERT FIGURE 6 ABOUT HERE]

[INSERT TABLE 3 ABOUT HERE]

We summarize each model's performance for a dataset using MAE, averaged over the posterior uncertainty. For example, for Dataset A the BG/BB has the best average out-of-sample prediction (MAE = 4.37, mean across replicates) closely followed by the HC (4.91). For Dataset B, the HMM (8.35) clearly out-performs the other three models, and Dataset C is best modeled by OF (7.09).

While those are posterior means of model performance, we also convey the degree of uncertainty in these assessments, using replicated datasets associated with the full posterior predictive distribution. To illustrate this uncertainty, we plot the predicted incremental sales for each posterior replicate for Dataset A and the observed daily incremental sales (Figure 7). By visual inspection of these tracking plots alone, it is difficult to detect whether the BG/BB truly predicts better than the other three models.

[INSERT FIGURE 7 ABOUT HERE]

While we would like to declare a single winning model for each dataset, the high level of uncertainty around the model predictions seems to raise a warning flag about making any strong statements about differences among the models. Therefore, we want to quantify the "shades of gray" in model performance by recognizing that when declaring a winning model the "vote" need not be unanimous.

Thus, instead of only examining posterior mean of MAE, we characterize its full distribution. For the highlighted Datasets A, B, and C, we show the distribution of each model's MAE across all replicates (Figure 8). Table 4 displays the corresponding distribution summaries (e.g., median and interquartile ranges of MAE across replicates).

[INSERT FIGURE 8 ABOUT HERE]

[INSERT TABLE 4 ABOUT HERE]

Not surprisingly, the densities of the performance measure of the four models are somewhat overlapping. For example, in Dataset A, while most of the mass of the BG/BB density is lower (better) than that of the HMM and HC densities, there is some probability that the HMM or HC has a lower MAE than the BG/BB. This suggests there is not a unanimous winner. By contrast, in Dataset C, for instance, there is much less overlap, suggesting that the OF has an even higher chance of a lower MAE than the others.

But what is the probability that each model is the winning model for a dataset? We take advantage of the Bayesian output to make this probability statement. Table 5 shows each model's winning percentage for Datasets A, B, and C. That winning percentage, or percent of votes, is the proportion of posterior worlds in which each model has the lowest error. For instance, the OF is quite clearly the winner for Dataset C, since it wins 81% of the time. For Dataset A, while the BG/BB is the winner, the distributions of the error for three of the four models overlap. So it is not surprising that they "split the votes," and the BG/BB wins 45% of the time compared to 27%, 6% and 22% for the HC, OF, and HMM, respectively.

[INSERT TABLE 5 ABOUT HERE]

### 5.3 Assessing the predictive value of the decision tree: in-sample

How accurate are the resulting recommendations from the tree? We answer this question to assess the tree's predictive value. First, we focus on the simple measure of the hit rate of the classification tree. The hit rate is the number of times the tree recommends a model that is in fact the best model to use on that dataset. Averaged across all models and iterations, the hit rate is 46%. We put this hit rate in context by noting that from a purely operational standpoint, the tree allows the analyst to run one model instead of four. In other words, by reducing the work of an analyst by 75%, the tree makes a recommendation of which model to use that is about twice as good as guessing (25% hit rate) randomly among the four models. This hit rate also fares well when compared to tougher comparative yardsticks, such as the proportional

chance criterion and maximum chance criterion (Morrison 1969), which yield benchmark hit rates of 28% and 35%, respectively. The latter metric is often hard to beat in a discriminant analysis setting. It assesses how much better our classifications are, compared to using the most common actual winner (in this case the HMM) every time. So the decision tree clearly offers some improvements over that simple (but often effective) approach.

However, this measure is purely an "in-sample" one: it uses the same 64 datasets for calibration and classification purposes, so it may be subject to overfitting. We describe a procedure next, random forests, that will allow us to reflect the uncertain nature of the tree itself and its application to holdout data.

### 5.4 Assessing the predictive value of the tree using random forests: out-of-sample

To answer the question about the value of the tree, using another "lens," we turn to another machine-learning method closely related to CART known as *random forests* (Breiman 2001a; Liaw and Wiener 2002). While the single classification tree we have described above takes into account the parameter and model uncertainty, it does not take into account uncertainty in the structure of the classification tree itself. The random forest captures extra variation around the classification. This requires many classification trees, so the random forest algorithm "grows" many trees (hence the "forest").

What is special about the random forest is that it has a built-in monitoring system to make sure it produces predictions that are likely to be validated on a holdout set and that are utilizing important predictor variables, both to prevent overfitting (Breiman 2001a; Liaw and Wiener 2002).

Fortunately, the random forest algorithm has a built in cross-validation procedure calculating an $n$-fold cross validation, where the holdout sample size, $n$, is typically about 1/3 of the cases (Breiman 2001a). The holdout misclassification rate, in the language of machine learning, is called the "out-of-bag" error rate, or the "generalized error rate," since it is intuitively similar to cross-validation error, which indicates the ability of the predictive model to generalize to cases outside of the given dataset.

The random forest out-of-sample error rates broken down by each model are in Table

6, and the hit rate across all four models is 48%. This closely matches the in-sample hit rate using one classification tree. It is encouraging to see that, even when a dataset is not used for calibration, it can be classified correctly with a high level of accuracy.

[INSERT TABLE 6 ABOUT HERE]

Looking more carefully at the classification tree and random forest results, several distinctive patterns arise. First, it is clear that the BG/BB and HMM models have substantially higher hit rates than the other two models. It seems that each of these polar opposite models (at least in terms of parameters and complexity) can serve as effective "representatives" to characterize the entire family of HMM models covered here.

This result raises the question about which of the two constraints/dimensions associated with our $2 \times 2$ framework is more important to capture - the presence of an "off" state or the existence of an absorbing state? A closer inspection of Table 6 clearly reveals the answer: classifying whether or not the data requires an absorbing-state model or a back-and-forth model is much more informative than the presence of an "off state." There is a high degree of confusion between the BG/BB and HC models, and likewise for HMM and OF, but relatively little confusion between BG/BB and OF, or HMM and HC. In Table 7 we aggregate the classifications across this single dimension, and see incredibly high hit rates (62% and 89%) when we ignore the presence/absence of the "off state."

[INSERT TABLE 7 ABOUT HERE]

While we have explored the predictive value of the decision tree, it is natural to ask what is driving its good predictive ability. In order to better understand the drivers of our strong classification capabilities, we now analyze the database characteristics' diagnostic value.

### 5.5 Which database characteristics are most diagnostic?

The output of the random forests uncovers which variables are most important in explaining classification success. This not only validates the decision tree obtained via CART, but it also quantifies variable importance. Variable importance in random forests is a measure of

the average improvement in prediction accuracy of a tree when this variable is included (and its values are intact) compared to when this variable's values are meaningless (arbitrarily permuted across observations).

Figure 9 displays each database characteristic's variable importance. This analysis confirms what we see in the classification tree: early trend, late trend, trend Gini (trend shape), and concentration are the four most important variables, and they are clearly separated out from the others. Among the four, however, late trend is the most important. This makes intuitive sense as the models differ in their abilities to generate decreasing or increasing patterns in aggregate sales over time. For instance, for a dataset with a strongly increasing late trend, the BG/BB and HC (due to their absorbing state) would not be able to capture it at all. This provides more evidence that, even before running any models, an analyst could use these easy-to-compute database characteristics to refine the decision about which model is likely to perform best.

[INSERT FIGURE 9 ABOUT HERE ]

### 5.6 How much value does the classification tree add?

What does the analyst gain by using our decision tree? From the above discussion, we find the decision tree nearly doubles the hit rate compared to uninformed guessing about which of the four models to run. And much of the remaining error rate is associated with the relatively unimportant distinction between the presence/absence of an "off" state.

But while this information helps ease the task of choosing the right model, it also tells us how well an analyst would do using the decision tree compared to running all four models for every dataset. So it is reasonable to ask: how much error would the analyst suffer by using only one model for all datasets? After all, this is the starting point for many analysts. Suppose the analyst only used the BG/BB for all datasets she encountered. How poorly would she have performed? We can compare the error incurred to the average performance if she always used the truly winning model for each dataset. Table 8 summarizes this analysis.

Running the BG/BB on all datasets yields an error 52% worse than using the true winning model (average MAE = 9.52 vs. 6.28). An analyst would do better by running only the HMM, which yields an error 21% worse than the using winning model (average MAE = 7.63).

By contrast, using the model recommended by decision tree for each dataset is the best option as it greatly reduces error to only 12% worse than the winning model (average MAE = 7.05). That is, in terms of relative error to the best model, only using the HMM is 75% worse than using the decision tree. So using the decision tree is a win-win, it requires 75% less effort and helps the analyst to avoid a 75% increase in relative error.

[INSERT TABLE 8 ABOUT HERE]

The value of the decision tree is even greater if we look beyond average performance and consider the worst case scenario of model performance. When examining the variability in performance, the 95% level of error for using any single model can be quite high. However, the decision tree greatly controls that upper tail of error. In particular, the high end of possible error for the HMM is 35% worse than the winner's error, but the decision tree is only 12% (Table 8).

In short, our tree shows that an analyst should not use the same model for all occasions, and clearly quantifies the cost of doing so.

## 6 General discussion and future directions

When researchers and managers regularly encounter a particular kind of data structure and regularly choose among a standard set of models, they often develop good intuition about when to use which model. Our approach rigorously quantifies and validates this kind of intuition through a well-structured decision tree.

For the case of a database of repeat purchases over time for a cohort of customers, we make specific recommendations about when to use the HMM and its constrained variants, and which dataset-level summaries are important for that decision. We find that for datasets exhibiting an early decreasing trend in aggregate sales, the BG/BB provides the best forecast when the trend continues to decrease even later in the calibration period. But when it looks like the trend has leveled off, the BG/BB frequently underpredicts and more complexity is often warranted. An interesting exception to this rule, is the case of high purchase concentration suggests that the "buy till you die" framework is still likely to provide the best forecast. This may be reflective of the customers exhibiting high heterogeneity in purchase and churn rates rather than a more complex back-and-forth state-switching process over time.

In the case of the $2 \times 2$ framework, the models are classified with strong evidence along one dimension (the presence of an absorbing state versus back-and-forth movement across states), but the data offers weaker evidence to help discriminate datasets and models along the other dimension (presence/absence of an "off state"). This may be surprising in light of many papers that add a "death" state to an HMM-like model. But it may be the case that such models work well mainly because of the constraint making that state absorbing and not necessarily because the behavior is "turned off" within it. This finding could have important implications for model builders and should be investigated more carefully in settings beyond this framework.

Beyond our HMM-based example, our proposed approach for empirical identification is more broadly relevant. We explicitly test the characteristics of datasets that distinguish one model from a related one. While this differs from a formal theoretical identification (e.g., using economic principles), it is aligned with the calls for such activities that have been arising more frequently in marketing (Hartmann et al. 2008).

The procedure that we propose is quite general: given the appropriate inputs (i.e., database characteristics), it can generate a decision tree prescribing which model should be used for any given dataset and any given outcome/goal of interest. Understanding the interplay between database characteristics and the relative performance of models (and model components) is a useful contribution beyond the illustrative (yet common) context presented here. While we illustrate it here with the HMM on forecasting incidence data (e.g., repeat purchasing of a cohort), it is agnostic to these choices. In general, the recipe for this method requires the following elements: (1) a consideration set of candidate models, (2) a set of predictor variables consisting of observed summary statistics from each dataset, and (3) the outcome variable, which is a choice of how to "pick the winner"; this requires a key managerial quantity and a loss function for computing the error measure.

Classification and regression trees and random forests, while popular in machine learning and statistics, are still relatively new to the field of marketing, so we hope our work will call more attention to this powerful and versatile tool. Furthermore, our application of it to the problem of model selection (as opposed to variable selection) is relatively uncommon even in the statistics literature, but it is clearly a natural and important issue in many marketing contexts.

Unlike traditional uses of classification methods, we add an extra twist by employing them in a fully Bayesian framework, allowing us to leverage the full posterior distribution. This differs from previous mixtures of Bayesian and classification methods, e.g. Bayesian CART (Chipman et al. 1998), since we construct a decision tree from information that already incorporates a joint posterior distribution. Our mix of Bayesian approaches with classification methods is a promising area of research for the interface of marketing, statistics, and machine learning. The combination of the two approaches represents an exciting blurring of methodological boundaries, and marketing problems like the one examined here have a great deal to offer in the debate between the "two cultures" of data modeling (statistics) and algorithmic modeling (machine learning) put forward by (Breiman 2001b).

As computational costs decrease and access to grid/cloud computing increases, the procedure we propose here will be even easier to do in a variety of contexts. Of course, one could argue that with greater computing power, there is less need to worry about selecting the single best model a priori – just run a bunch of models and pick the best one. But this logic is flawed for several reasons. First, our analysis focuses on performance in a holdout period, not in-sample fit. Second, and related, there is great danger in choosing models that are overly complex and excessively customized to every different dataset. And third, we believe strongly in exploring and learning from the underlying patterns that are driving the observed data patterns. This kind of "data science" will not only help analysts create and choose better models, but will help managers make better tactical decisions to order to create and extract more value from their customer relationships.

## References


Bladt, Mogens, M. F. Neuts. 2003. Matrix-Exponential Distributions: Calculus and Interpretations Via Flows. *Stochastic Models* **19**(1) 113–124.

Breiman, Leo. 2001a. Random Forests. *Machine Learning* **45** 5–32.

Breiman, Leo. 2001b. Statistical Modeling: The Two Cultures. *Statistical Science* **16**(3) 199–231.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, Charles J. Stone. 1984. *Classification and Regression Trees*. Chapman & Hall, New York, NY.

Chib, Siddhartha. 1995. Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association* **90**(432) 1313–1321.

Chipman, Hugh A., Edward I. George, Robert E. McCulloch. 1998. Bayesian CART Model Search. *Journal of the American Statistical Association* **93**(443) 935–948.

Cooper, Lee G., Masao Nakanishi. 1988. *Market-Share Analysis: Evaluating Competitive Marketing Effectiveness*. Kluwer Academic Publishers.

Donkers, Bas, Peter C. Verhoef, Martijn de Jong. 2007. Modeling CLV: A Test of Competing Models in the Insurance Industry. *Quantitative Marketing and Economics* **5**(2) 163–190.

Everson, Philip J., Eric T. Bradlow. 2002. Bayesian Inference for the Beta-Binomial Distribution via Polynomial Expansions. *Journal of Computational and Graphical Statistics* .

Fader, Peter S., Bruce G. S. Hardie, C. Y. Huang. 2004. A Dynamic Changepoint Model for New Product Sales Forecasting. *Marketing Science* **23**(1) 59–65.

Fader, Peter S., Bruce G. S. Hardie, Jen Shang. 2010. Customer-Base Analysis in a Discrete-Time Noncontractual Setting. *Marketing Science* **29**(6) 1086–1108.

Gelman, Andrew, John B. Carlin, Hal S. Stern, Donald B. Rubin. 2004. *Bayesian Data Analysis*. 2nd ed. Chapman & Hall, New York, NY.

Gelman, Andrew, Donald B. Rubin. 1992. Inferences from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7**(4) 457–472.

Gelman, Andrew, Donald B. Rubin. 1995. Avoiding Model Selection in Bayesian Social Research. *Sociological Methodology* **25** 165–173.

Geweke, John. 1992. Evaluating the Accuracy of Sampling-based Approaches to Calculating Posterior Moments. J. M. Bernado, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds., *Bayesian Statistics 4*. Oxford University Press.

Hanssens, Dominique M., , Peter S. H. Leeflang, Dick R. Wittink. 2005. Market Response Models and Marketing Practice. *Applied Stochastic Models in Business and Industry* **21**(4/5) 423–434.

Hartmann, Wes, Puneet Manchanda, Harikesh Nair, Matt Bothner, Peter Dodds, Dave Godes, Karthik Hosanagar, Catherine Tucker. 2008. Modeling Social Interactions: Identification, Empirical Methods and Policy Implications. *Marketing Letters* **19**(4) 287–304.

Kamakura, Wagner A., Gary J. Russell. 1989. A Probabilistic Choice Model for Market Segmentation and Elasticity Structuring. *Journal of Marketing Research* **26**(4) 379–90.

Lenk, Peter J. 2009. Simulation Pseudo-Bias Correction to the Harmonic Mean Estimator of Integrated Likelihoods. *Journal of Computational and Graphical Statistics* **18**(4) 941–960.

Liaw, Andy, Matthew Wiener. 2002. Classification and Regression by randomForest. *R News* **2**(3) 18–22.

Liechty, John C., Rik Pieters, Michel Wedel. 2003. Global and Local Covert Visual Attention: Evidence from a Bayesian Hidden Markov Model. *Psychometrika* **68**(4) 519–541.

Ma, Shaohui, Joachim Buschken. 2011. Counting Your Customers from an "Always a Share" Perspective. *Marketing Letters* **22**(3) 243–257.

Montgomery, Alan L., Shibo Li, Kannan Srinivasan, John C. Liechty. 2004. Modeling Online Browsing and Path Analysis Using Clickstream Data. *Marketing Science* **23**(4) 579–595.

Montoya, Ricardo, Oded Netzer, Kamel Jedidi. 2010. Dynamic Allocation of Pharmaceutical Detailing and Sampling. *Marketing Science* **29**(5) 909–924.

Morrison, Donald G. 1969. On the Interpretation of Discriminant Analysis. *Journal of Marketing Research* **6**(2) 156–163.

Netzer, Oded, James M. Lattin, V. Srinivasan. 2008. A Hidden Markov Model of Customer Relationship Dynamics. *Marketing Science* **27**(2) 185–204.

Newton, Michael A., Adrian E. Raftery. 1994. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**(1) 3–48.

O'Cinneide, Colm A. 1990. Characterization of Phase-Type Distributions. *Communications in Statistics: Stochastic Models* **6**(1) 1–57.

Plummer, Martyn, Nicky Best, Kate Cowles, Karen Vines. 2006. CODA: Convergence Diagnosis and Output Analysis for MCMC . *R News* **6**(1) 7–11.

Raftery, Adrian E., S. M. Lewis. 1992. How Many Iterations in the Gibbs Sampler? J. M. Bernado, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds., *Bayesian Statistics 4*. Oxford University Press.

Sabavala, Darius J., Donald G. Morrison. 1977. A Model of TV Show Loyalty. *Journal of Advertising Research* **17**(6) 35–43.

Schmittlein, David C., Donald G.Morrison, Richard A. Colombo. 1987. Counting Your Customers: Who Are They and What Will They Do Next? *Management Science* **33**(1) 1–24.

Schweidel, David A., Eric T. Bradlow, Peter S. Fader. 2011. Portfolio Dynamics for Customers of a Multi-Service Provider. *Management Science* **57**(3).

Schweidel, David A., Peter S. Fader. 2009. Dynamic Changepoints Revisited: An Evolving Process Model of New Product Sales. *International Journal of Research in Marketing* **26**(2) 119–124.

Schweidel, David A., Peter S. Fader, Eric T. Bradlow. 2008. Understanding Subscriber Retention Within and Across Cohorts Using Limited Information. *Journal of Marketing* **72**(1) 82–94.

Spiegelhalter, David J., Nicola G. Best, Bradley P. Carlin, Angelika Van Der Linde. 2002. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society, Series B* **64**(4) 583–639.

Stephens, Matthew. 2000. Dealing with Label-Switching in Mixture Models. *Journal of the Royal Statistical Society Series B* **62** 795–809.

Tanner, Martin A., Wing H. Wong. 1987. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association* **82**(398).

Wittink, Rick R., M. J. Addona, W. J. Hawkes, J.C. Porter. 1988. SCAN*PRO: The Estimation, Validation and Use of Promotional Effects Based on Scanner Data. *Internal Paper* Cornell University.

Zheng, Zhiqiang (Eric), Peter Fader, Balaji Padmanabhan. 2011. From Business Intelligence to Competitive Intelligence: Inferring Competitive Measures Using Augmented Site-Centric Data. *Information Systems Research* **23**(3) 698–720.

**Tables and Figures**

| | |
|---|---|
| Frequency | How many active days of transactions are there per customer? |
| Penetration | How many unique customers have made at least one transaction? |
| Concentration | How is activity spread out among customers (i.e., what fraction of all transactions was made by the top 20% of customers)? |
| Top 5% Level | How much are the most active customers purchasing (i.e., what level of transactions is the cutoff for the top 5% of most active customers)? |
| Early Trend | What is the trend in the *first half* of the calibration period (i.e., drop from first to middle day as a percentage of the size of the customer base)? |
| Late Trend | What is the trend in the *second half* of the calibration period (i.e., drop from middle to last day as a percentage of the size of the customer base)? |
| Trend Gini | How much does the actual curve deviate from a line connecting the first and last days (i.e., how much area is there below the trend line and the curve, as percentage of the levels of the line, *a la* the Gini coefficient)? |
| Trend Variability | How much day-to-day variation is present in the calibration period (i.e., standard deviation of incremental sales)? |

Table 1: *Database characteristics capture features of a longitudinal incidence dataset and can be computed from summary plots (e.g., histogram and tracking plot).*

| Dataset | A | B | C |
|---|---|---|---|
| Frequency | 11% | 12% | 11% |
| Penetration | 48% | 83% | 68% |
| Concentration | 84% | 55% | 63% |
| Top 5% Level | 63% | 40% | 43% |
| Early Trend | -17% | -29% | -7% |
| Late Trend | -3% | 1% | 0% |
| Trend Shape | 27% | 46% | 13% |
| Trend Variability | 4% | 6% | 2% |

Table 2: *Database characteristics for the three highlighted datasets.*

| Posterior Mean of MAE | | | |
|---|---|---|---|
| | A | B | C |
| BGBB | **4.37** | 9.71 | 14.13 |
| HC | 4.91 | 10.16 | 17.29 |
| OF | 6.52 | 10.20 | **7.09** |
| HMM | 5.13 | **8.35** | 9.44 |

Table 3: *For each dataset and each model's posterior draw, the out-of-sample forecast is generated from the posterior predictive distribution and the mean absolute error (MAE) is computed. The posterior mean of the MAE values for each model-dataset pair is shown here. The lowest error value (i.e., winning model) for each dataset is in bold.*

| Posterior Distribution of MAE | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Dataset A | | | Dataset B | | | Dataset C | | |
| | 25% | 50% | 75% | 25% | 50% | 75% | 25% | 50% | 75% |
| BGBB | **3.7** | **4.2** | **4.8** | 8.3 | 9.6 | 11.0 | 11.9 | 14.0 | 16.2 |
| HC | 4.0 | 4.6 | 5.5 | 8.9 | 10.1 | 11.4 | 15.4 | 17.3 | 19.1 |
| OF | 4.9 | 6.2 | 7.8 | 8.6 | 10.0 | 11.6 | **6.3** | **7.0** | **7.8** |
| HMM | 4.1 | 4.7 | 5.8 | **7.0** | **8.1** | **9.5** | 7.7 | 9.0 | 10.7 |

Table 4: *For each dataset and each model's posterior draw, the out-of-sample forecast is generated from the posterior predictive distribution. Mean absolute error (MAE) is computed for each replicate dataset. The posterior quantiles (25%, 50%, and 75%) of the values across for each dataset are shown here. The lowest value (i.e., winning model) for each dataset is in bold.*

| Posterior Probability of Model Winning | | | | |
|---|---|---|---|---|
| Dataset | BGBB | HC | OF | HMM |
| A | 45% | 27% | 6% | 22% |
| B | 20% | 11% | 16% | 53% |
| C | 1% | 0% | 81% | 18% |

Table 5: *The posterior probabilities of each model "winning" are computed as the proportion of replicated datasets (e.g., "MCMC worlds") in which each model has the lowest MAE. These winning percentages illustrate the uncertainty in declaring a winner.*

| | BGBB | HC | HMM | OF | Hit Rate |
|---|---|---|---|---|---|
| BGBB | 840 | 151 | 215 | 114 | 64% |
| HC | 372 | 247 | 145 | 154 | 27% |
| HMM | 430 | 40 | 1056 | 677 | 48% |
| OF | 259 | 7 | 742 | 951 | 49% |

Table 6: *This table shows the out-of-sample classification for each of the 6,400 cases (dataset-world pairs) from the random forest. Rows indicate which model actually fit the data best. Columns indicate which model was recommended by the random forest's classification for that dataset using a 2/3 sample for calibration (in-sample) and 1/3 sample for validation (out-of-sample). The hit rate is the proportion of each type of dataset correctly classified out-of-sample (i.e., the diagonal entries divided by row sums).*

| | Absorbing | Back-and-Forth | Hit Rate |
|---|---|---|---|
| Absorbing | 1383 | 855 | 62% |
| Back-and-Forth | 466 | 3696 | 89% |

Table 7: *This table combines cases of datasets and classifications into models with Absorbing states (BG/BB and HC) and Back-and-Forth transitions (OF and HMM). That is, by ignoring the presence/absense of an "off"/"death" state, the hit rates are quite high. Like Table 6, these are out-of-sample classifications, so the hit rate is the proportion of cases correctly classified.*

| | BGBB | HC | OF | HMM | Tree | Winner |
|---|---|---|---|---|---|---|
| Mean MAE | 9.52 | 9.19 | 8.14 | 7.63 | 7.05 | 6.28 |
| % Worse than Winner | 52% | 46% | 30% | 21% | 12% | – |
| 95% MAE | 22.34 | 18.97 | 16.70 | 15.63 | 13.00 | 11.57 |
| % Worse than Winner | 93% | 64% | 44% | 35% | 12% | – |

Table 8: *This table shows the absolute and relative benefits of using the classification tree over running each model for all datasets. The MAE values reflect performance of running each model for all datasets compared to following the tree's recommendation (Tree) and always selecting the model with best out-of-sample fit for each dataset (Winner). The percentages illustrate the loss compared to the best fitting model (Winner).*

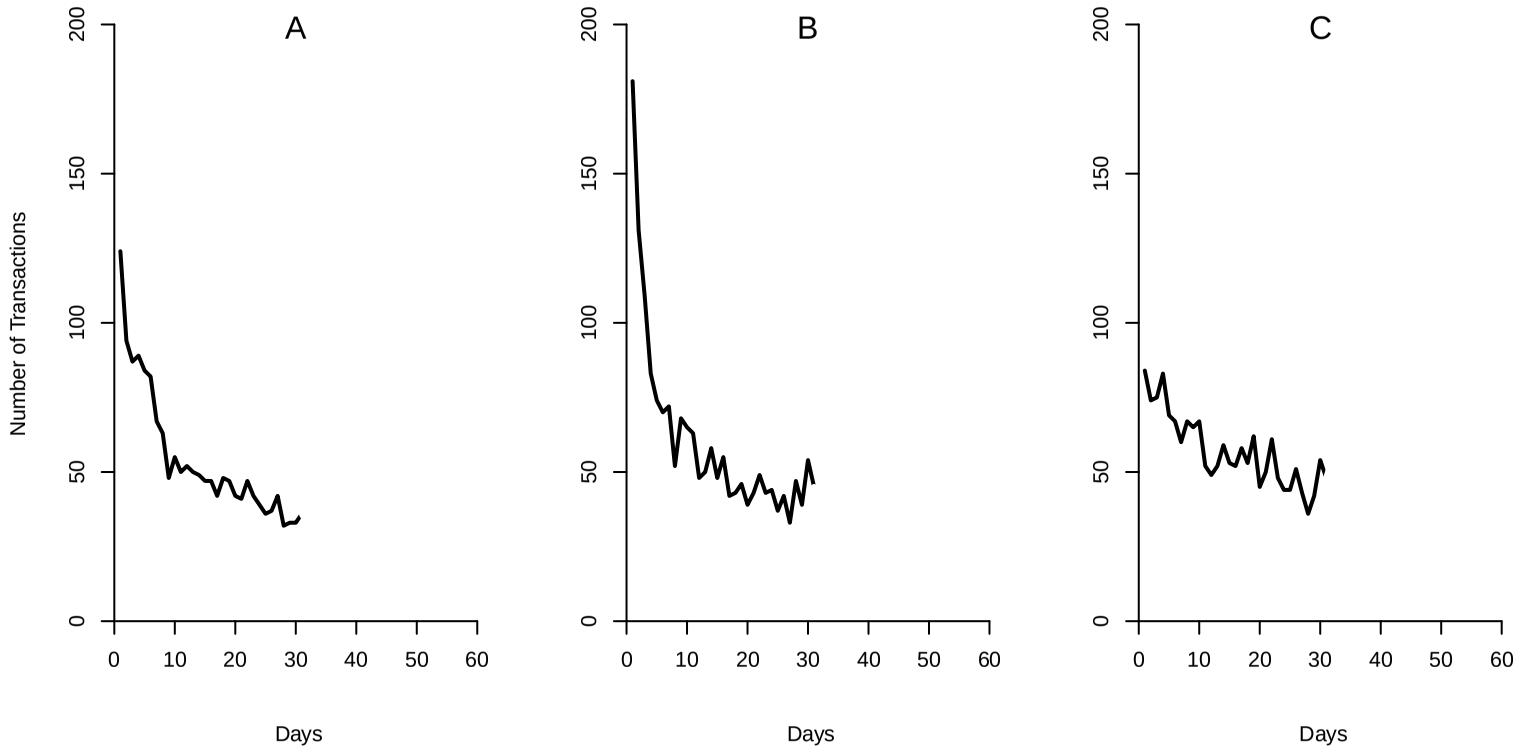


Figure 1: *These plots show three different commonly observed patterns of aggregate sales over time (e.g., arising from a product launch or repeat purchasing of cohort of new customers).*

| | State 2 is Cold | State 2 is Off |
|---|---|---|
| Backwards and Forwards Transitions | HMM<br>$\mathbf{p}_i = (p_{1i}, p_{2i})$ and $\mathbf{\Theta}_i = \begin{pmatrix} 1-\theta_{12i} & \theta_{12i} \\ \theta_{21i} & 1-\theta_{21i} \end{pmatrix}$ | On And Off<br>$\mathbf{p}_i = (p_{1i}, 0)$ and $\mathbf{\Theta}_i = \begin{pmatrix} 1-\theta_{12i} & \theta_{12i} \\ \theta_{21i} & 1-\theta_{21i} \end{pmatrix}$ |
| Only Forward Transitions | Hot Then Cold<br>$\mathbf{p}_i = (p_{1i}, p_{2i})$ and $\mathbf{\Theta}_i = \begin{pmatrix} 1-\theta_{12i} & \theta_{12i} \\ 0 & 1 \end{pmatrix}$ | BGBB<br>$\mathbf{p}_i = (p_{1i}, 0)$ and $\mathbf{\Theta}_i = \begin{pmatrix} 1-\theta_{12i} & \theta_{12i} \\ 0 & 1 \end{pmatrix}$ |

Figure 2: *This $2 \times 2$ table summarizes the nested model relationships among the HMM and its variants. The rows and columns illustrate how constraints on two model components lead to different models. The $\mathbf{p}_i$ and $\mathbf{\Theta}_i$ are the individual-level within-state transaction probabilities and between-state transition probability matrix, respectively.*

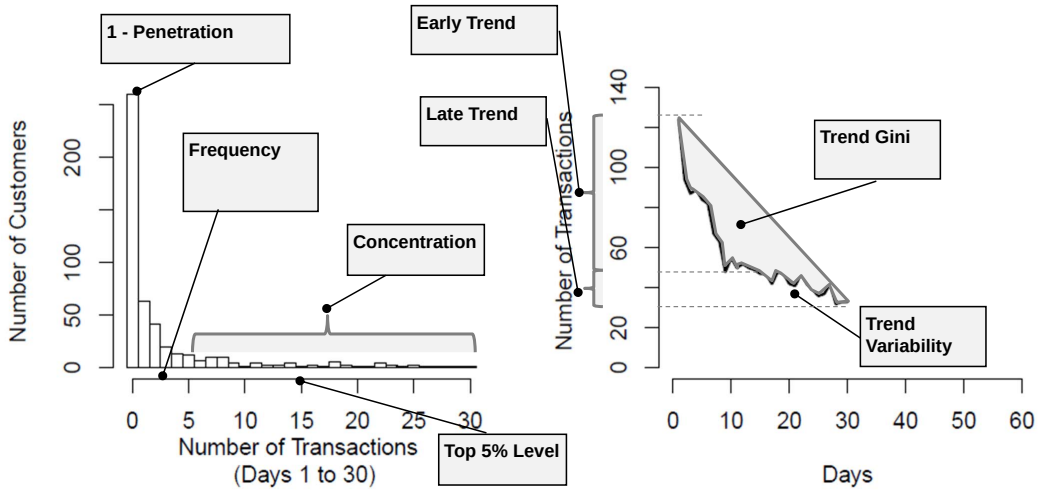


Figure 3: *The figure and annotations illustrate how the observed summary statistics arise naturally from plots that managers typically examine when deciding which model(s) to run. The histogram (left) shows how the number of transactions varies across customers in the observation period, while the tracking plot (right) shows incremental transactions of the cohort over the same period.*

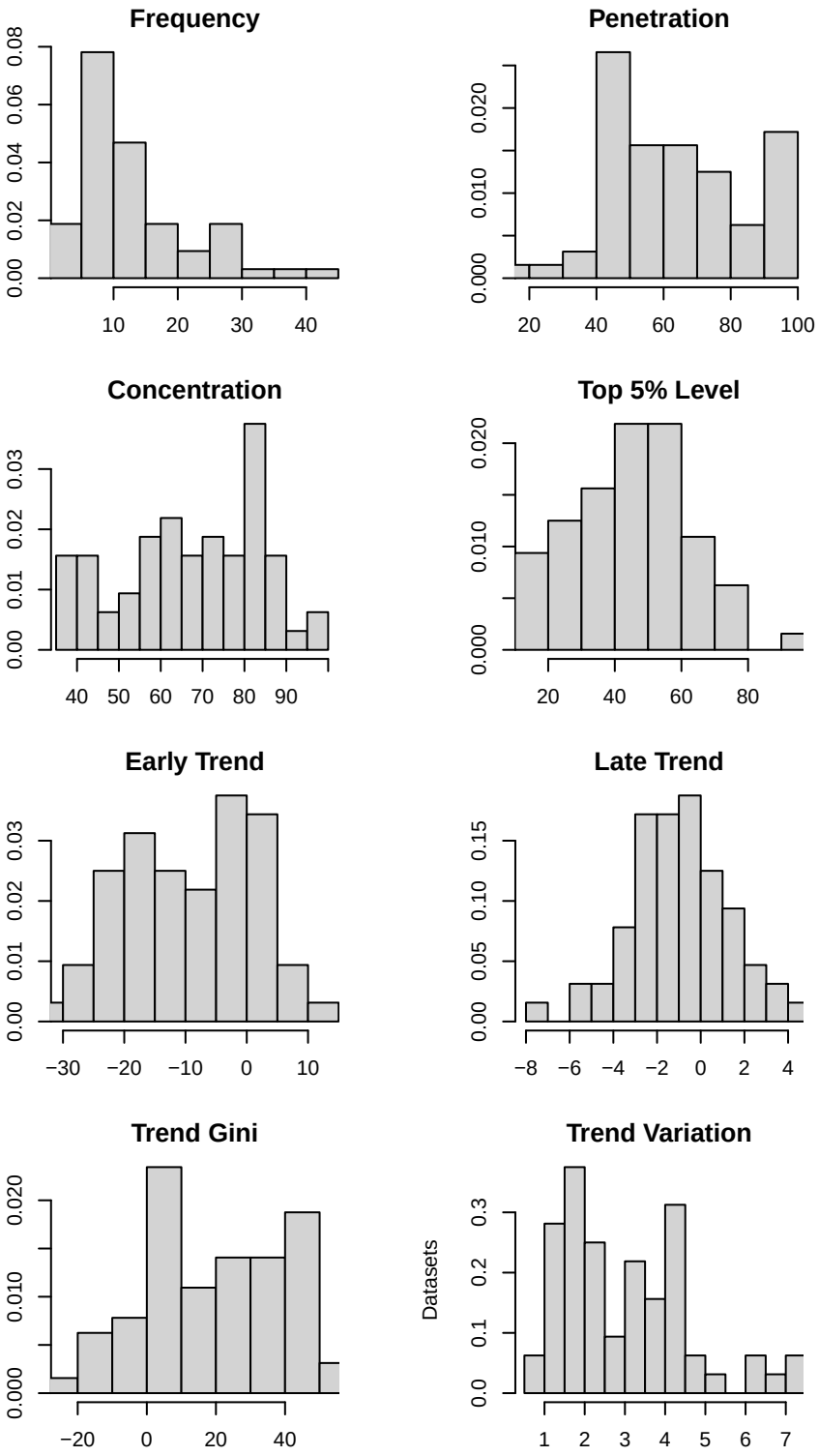


Figure 4: *These histograms summarize the variability for each database characteristics across all 64 datasets.*

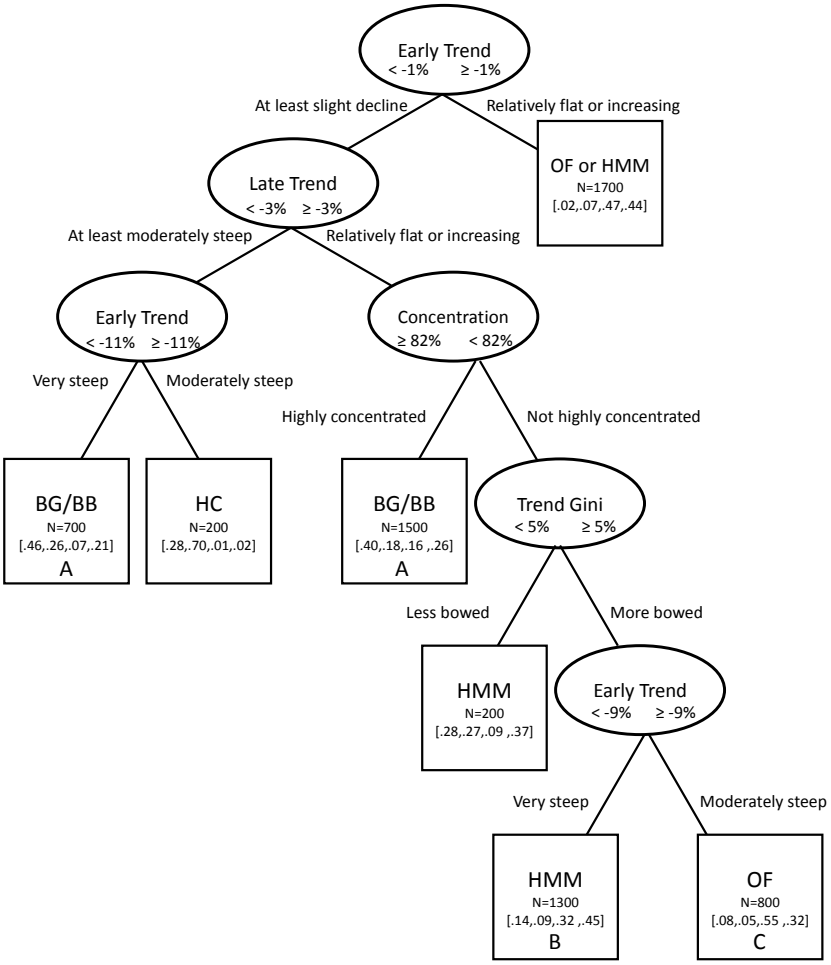


Figure 5: *The classification tree is estimated on 6,400 cases, using 100 posterior samples for each of the 64 datasets. For any dataset, the tree should be read from top to bottom: the ovals represent the partitions, the rectangles indicate the terminal nodes, and the listed model is the recommended one for that particular combination of database characteristics. Also listed is a vector summarizing the posterior winning percentage for all four models (left to right: BG/BB, HC, OF, HMM). The highlighted Datasets A,B, and C appear where they are best classified.*

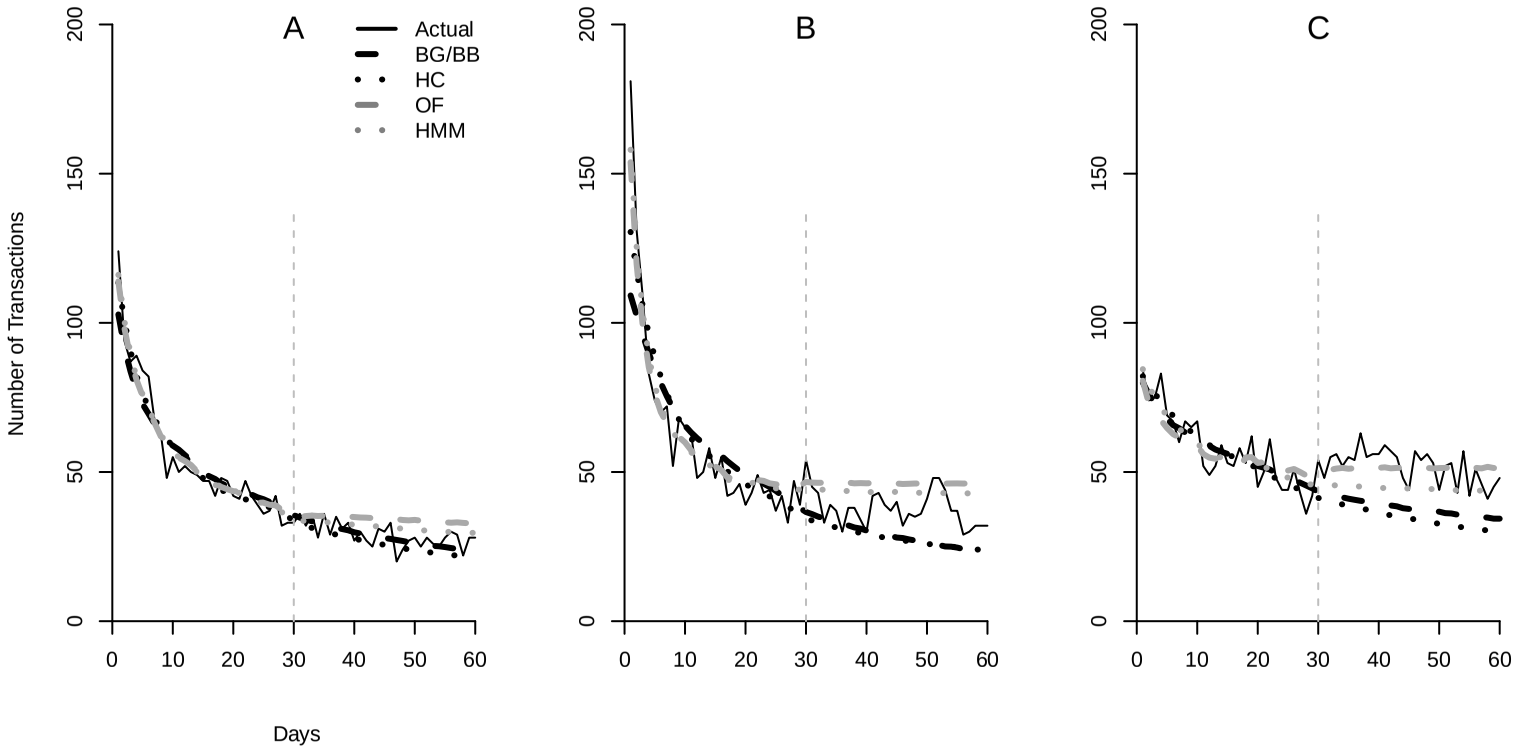


Figure 6: *The three example datasets are shown with the predictions arising from each of the four models. The mean of each model's posterior predictive distribution is shown.*

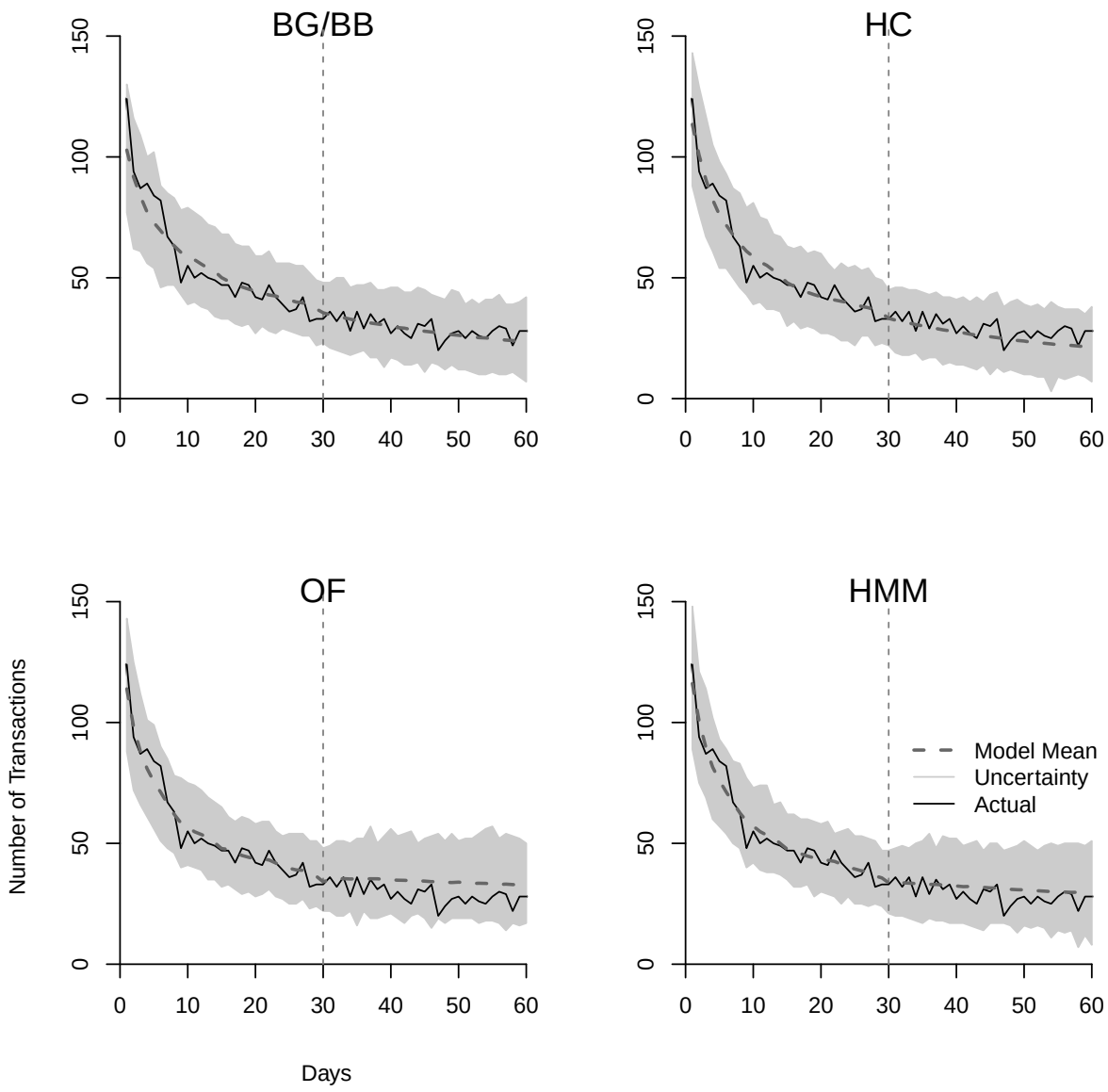


Figure 7: *The four plots illustrate the range of variability in model prediction for Dataset A. The observed data (solid line) is fit by each model. Each model's posterior predictive mean (dashed line) and its full distribution from 1,000 posterior draws (light gray shading) are displayed.*

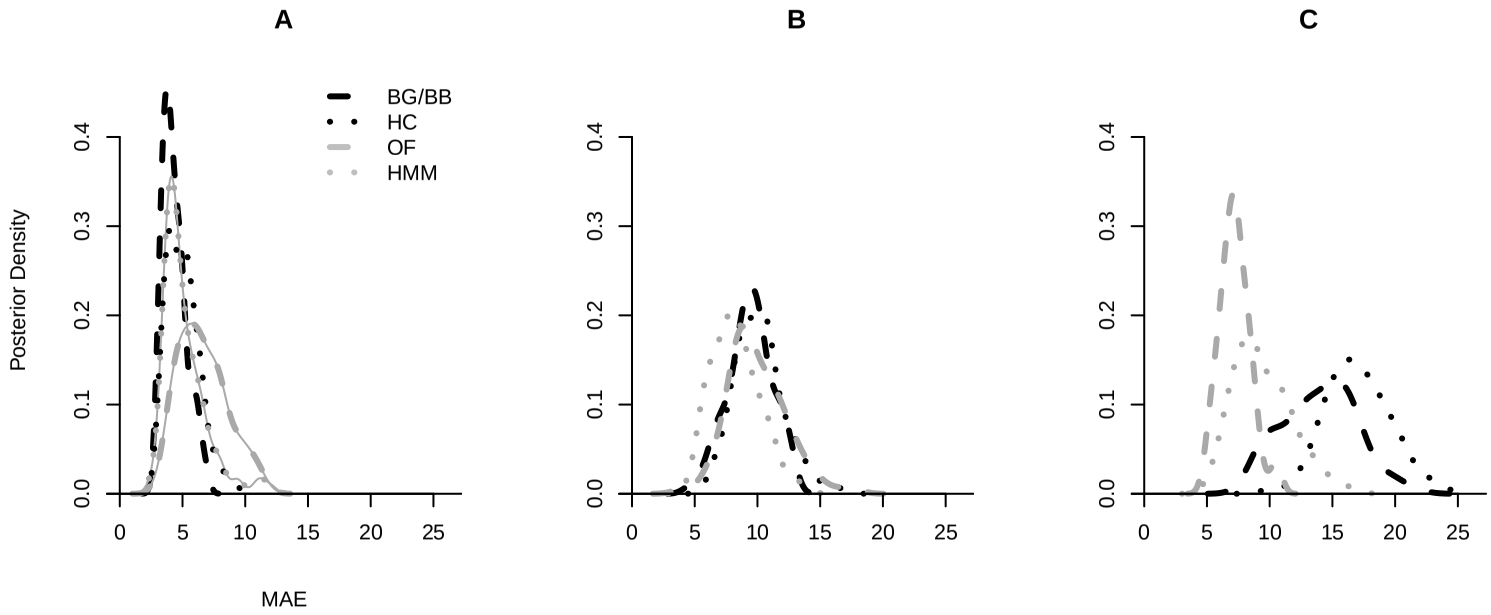


Figure 8: *These figures show the posterior distribution of out-of-sample mean absolute error (MAE) for the three highlighted datasets. These are densities of error measures, so smaller values indicate better model fit. When the densities overlap a great deal (e.g., Dataset A), there is not a clear winning model. When one density that stands out as better than the others (e.g., dataset C), there is a clear winning model.*

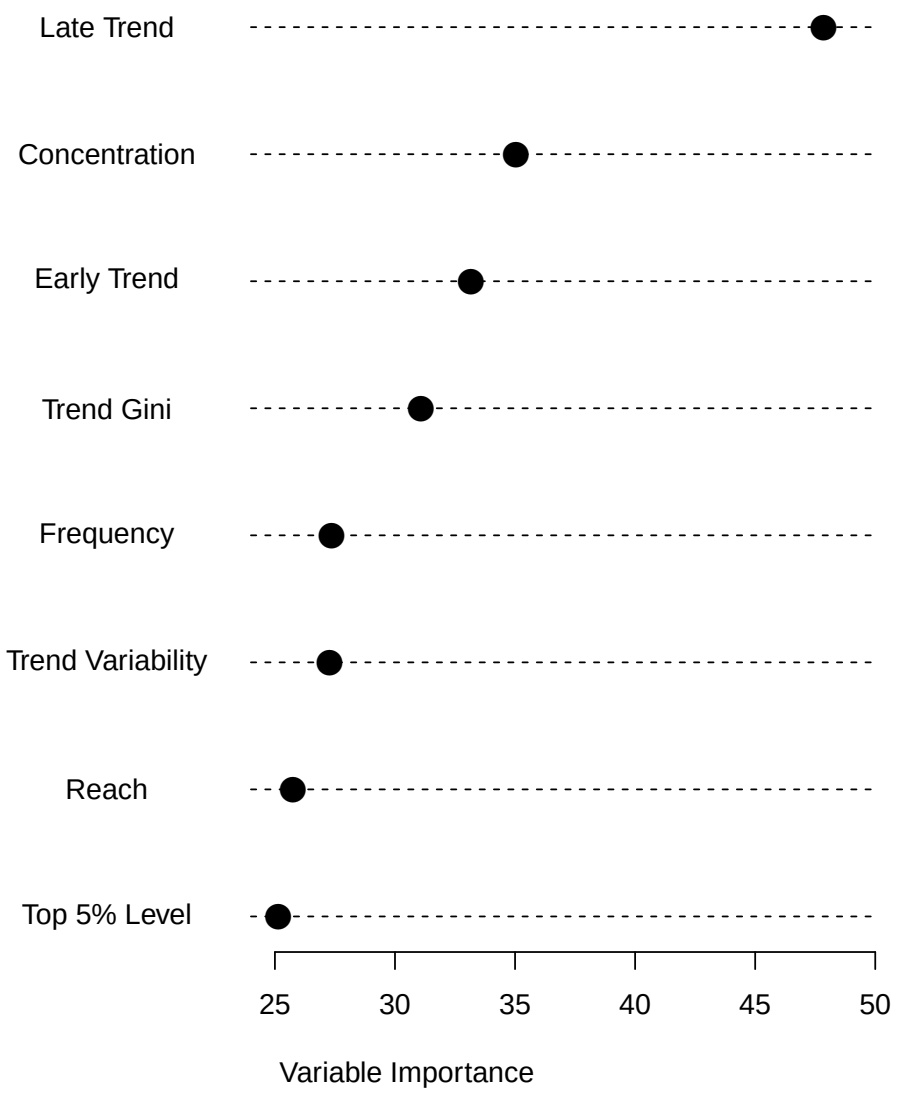


Figure 9: *This plot shows the relative importance of each database characteristic (predictor variable) used in the classification trees obtained from the random forest. The most important variables are those that provide the largest increase in out-of-sample (out-of-bag) classification hit rate, averaged across all trees in the forest. The late trend, early trend, and concentration are clearly the three most important, and confirmed by appearing in the classification tree obtained via CART methods. The next most important variable is trend Gini (trend shape), which also appears in the classification tree.*

## Appendix 1. Hierarchical Bayes sampler details

We provide the computational details for the models that we ran. We provide the details of the sampler for the general HMM with $S$ states. It can be constrained for the two-state HMM and each of its nested models, as described in Section 2.

The MCMC procedure generates draws from the joint posterior

$$[\mathbf{Z}_{it}, \mathbf{p}_i, \Theta_i, \mathbf{a}_p, \mathbf{b}_p, \alpha, \pi|\mathbf{Y}] = \prod_{i=1}^{I}\prod_{t=1}^{T}[Y_{it}|Z_{it}, p_i][Z_{it}|Z_{i,t-1}, \Theta_i, \pi] \\ [p_i|\mathbf{a}_p, \mathbf{b}_p, \mathbf{Z}_i, \mathbf{Y}_i][\Theta_i|\alpha, \mathbf{Z}_i][\mathbf{a}_p, \mathbf{b}_p][\alpha][\pi]$$

with constants ($I$ individuals and $T$ time periods), individual-level parameters ($\mathbf{p}_i$ and $\Theta_i$), and population-level parameters ($\mathbf{a}_p$, $\mathbf{b}_p$, $\alpha$, and $\pi$).

The procedure obtains these draws by alternating between the following conditional distributions:

$$\begin{aligned} &[\mathbf{Z}_i|\mathbf{Y}_i, \Theta_i, \mathbf{p}_i, \pi] \\ &[\mathbf{p}_i|\mathbf{Y}_i, \mathbf{Z}_i, \mathbf{a}_p, \mathbf{b}_p] \\ &[\Theta_i|\mathbf{Z}_i, \alpha] \\ &[\mathbf{a}_p, \mathbf{b}_p|\mathbf{p}] \\ &[\alpha|\mathbf{\Theta}] \\ &[\pi|\mathbf{Z}] \end{aligned} \tag{8}$$

For each entry in the "dataset of datasets" described in Section 3, we estimate all four models from the $2\times2$ framework. We use 64 datasets, 4 models per dataset, 2 chains per model, resulting in 512 independent MCMC chains. We run each chain for at least 50,000 iterations depending on the convergence criterion for that given model and chain. This requires over 1,000 days of computing time on a single core. Instead of using only one core, we distributed the computational task to take advantage of the parallel structure of the task. On the Amazon Elastic Computing Cloud, we used 64 nodes for 48 hours ("node-hours" ), where each node contains 8 cores (thanks to a grant from Amazon Web Services). We ran each MCMC chain on one of 512 cores, so we finished running all the models in just two days.

Each model is estimated using a version of the MCMC sampler for the HMM with certain components shut off or not. The code is available upon request. For each chain and for each pair of chains for each model, we perform a set of within-chain diagnostics for convergence and computation of effective sample size, and also across-chain diagnostics for post-convergence mixing – all recommended now as standard practice (Gelman et al. 2004; Gelman and Rubin 1992; Geweke 1992; Plummer et al. 2006; Raftery and Lewis 1992).

The draws of model parameters, $\mathbf{\Omega}$, and latent states, $\mathbf{Z}^*$, have been obtained using a data-augmented Gibbs sampler (Tanner and Wong 1987) with an embedded Metropolis-Hastings step. Below we describe how each subset of parameters was drawn from its corresponding conditional distribution in the MCMC procedures.

1. Generate $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iT})$. The customer's latent-state sequence is drawn via the forward-backward algorithm. The latent states are sampled starting at $t = T$ moving backwards based on the probabilities defined recursively starting at $t = 1$ moving forwards using dynamic programming. For the case of $S = 2$, given the observed outcome at $t$ and the probability of being in either state at $t - 1$, each element $\delta_{i,t,k}$ is a sum of

the two elements from $t-1$ weighted by the probability of the corresponding transition probabilities. Then the probability of drawing state $k$ is

$$\begin{aligned}
[\mathbf{Z}_{it}|\mathbf{Y}_i, \Theta_i, \mathbf{p}_i, \pi] &= \frac{\delta_{i,t,k}}{\delta_{i,t,1} + \ldots + \delta_{i,t,S}}, \\
\delta_{i,t,1} &= p_{1i}^{y_{it}}(1-p_{1i})^{1-y_{it}}\left(\delta_{i,t-1,1}\cdot\theta_{11i} + \ldots + \delta_{i,t-1,S}\cdot\theta_{S1i}\right) \\
&\ldots \\
\delta_{i,t,S} &= p_{Si}^{y_{it}}(1-p_{Si})^{1-y_{it}}\left(\delta_{i,t-1,1}\cdot\theta_{1Si} + \ldots + \delta_{i,t-1,S}\cdot\theta_{SSi}\right).
\end{aligned} \tag{9}$$

Once the sequences from $1, \ldots, T$ are drawn for all individuals, then conditioning on those sampled latent states, as if they were data (i.e., data augmentation), simplifies the subsequent conditional distributions. Hence, we define a vector, $\mathbf{N}_i$, where each element counts the number of times an individual spent a day in each latent state, $N_{ij} = \sum_{t=1}^{T} 1\{Z_{i,t-1} = j\}$. We also define a matrix, $\mathbf{N}_{it}$, where each entry $j,k$ counts the number of transitions made between each pair of latent states, $N_{ijk} = \sum_{t=1}^{T} 1\{Z_{i,t-1} = j\}1\{Z_{it} = k\}$.

2. Generate $\mathbf{p}_i = (p_{1i}, \ldots, p_{Si})$. The customer's purchase probability vector is sampled directly from a beta distribution. The use of independent beta priors for each probability yields a beta posterior distributions (since the likelihoods have no covariates). For state $k$, the prior and posterior are

$$\begin{aligned}
[p_{ki}] &= \text{beta}(a_{pk}, b_{pk}) \\
[p_{ki}|\mathbf{Y}_i, \mathbf{Z}_i, \mu_{\mathbf{p}}, \phi_{\mathbf{p}}] &= \text{beta}\left(a_{pk} + \sum_{t=1}^{T} y_{it}1\{Z_{it} = k\}, b_{pk} + N_{ij} - \sum_{t=1}^{T} y_{it}1\{Z_{it} = k\}\right),
\end{aligned} \tag{10}$$

where $a_{pk}$ and $b_{pk}$ are the shape parameters of the beta distribution, and $\mu_{pk}$ and $\phi_{pk}$ are the mean and polarization index, as defined in Section 2. To ensure $p_{1i} \geq p_{2i}$, we use rejection sampling of the whole vector.

3. Generate $\Theta_i$. The customer's transition probability matrix must have its rows sum to 1, so it is a multinomial vector. Using independent Dirichlet priors on each row yields Dirichlet posteriors. For the probability of moving from $j$ to any other state $1, \ldots, S$, the prior and posterior are

$$\begin{aligned}
[\theta_{j,1:S,i}] &= \text{Dirichlet}(\alpha_{j,1:S}) \\
[\theta_{j,1:S,i}|\mathbf{Z}_i, \mu_{\theta j}, \phi_{\theta j}] &= \text{Dirichlet}\left((\alpha_{j,1} + N_{ij,1}, \ldots, \alpha_{j,S} + N_{ij,S})\right),
\end{aligned} \tag{11}$$

where the $1:S$ indexes a vector of parameters and where $\mathbf{r}_j$ is the vector of Dirichlet shape parameters, which can be summarized by mean probability vector, $\mu_{\theta j} = \alpha_j / \left[\sum_{k=1}^{S} \alpha_{jk}\right]$, and polarization index $\phi_{\theta j} = 1/\left[1 + \sum_{k=1}^{S} \alpha_{jk}\right]$.

4. Generate $\pi$. The initial latent-state membership probability vector depends on the latent states across all individuals at $t = 1$. Defining $N1_k = \sum_{i=1}^{I} 1\{Z_{k1} = k\}$, the uniform

hyperprior and the posterior are,

$$\begin{aligned}[\pi] &= \text{Dirichlet}(1,\ldots,1) \\ [\pi|\mathbf{Z}] &= \text{Dirichlet}\,(1+N1_1,\ldots,1+N1_S)\end{aligned} \tag{12}$$

5. Generate $(a_p, b_p)$. There is a highly uninformative hyperprior for each shape parameter of each beta distribution characterizing the heterogeneity of state-specific purchase propensities. For state $k$, the hyperprior and posterior are

$$\begin{aligned}[a_{pk}, b_{pk}] &\propto (a_{pk}+b_{pk})^{-5/2} \\ [\mathbf{a}_{pk}, \mathbf{b}_{pk}|\mathbf{p}_k] &\propto L_{\text{beta}}(p_{ki},\ldots,p_{kI})(a_{pk}+b_{pk})^{-5/2},\end{aligned} \tag{13}$$

where $L_{\text{beta}}$ is the beta density function, and the prior distribution proportional to $(a+b)^{-5/2}$ is recommended by Gelman et al. (2004). That prior is uniform on the beta distribution $a/(a+b)$ and considered weakly informative on the polarization index $(1+a+b)^{-1}$. Since the posterior has no closed form expression, we use a Metropolis-Hastings step with a log-Normal proposal density. Its tuning parameter, or variance, is set to $0.05$ to obtain an appropriate acceptance probability.

6. Generate $\alpha$. The Dirichlet distribution shape parameters $\alpha_j = (\alpha_{j1},\ldots,\alpha_{jS})$ are generated by a generalization of the procedure used to generate the shape parameters of the beta distributions. For state $j$ (i.e., row $j$ of the transition probability matrix), the hyperprior and posterior are

$$\begin{aligned}[\alpha_j] &= (\sum_{k=1}^{S}\alpha_{jk})^{-(2S+1)/2} \\ [\alpha_j|\theta_{j,1:S}] &\propto L_{\text{Dirichlet}}(\theta_{j,1:S,i},\ldots\theta_{j,1:S,I})L_{\text{gamma}}(\alpha_j),\end{aligned} \tag{14}$$

where $L_{\text{Dirichlet}}$ is the Dirichlet density function. Again, the prior distribution proportional to $(\alpha_1+\ldots+\alpha_S)^{-(2S+1)/2}$ is a generalization for the Dirichlet shape parameters of the prior used for the beta shape parameters (Everson and Bradlow 2002). Since the posterior has no closed form expression, we use a Metropolis-Hastings step with a log-Normal proposal density. Its tuning parameter, or variance, is set to $0.05$ to obtain an appropriate acceptance probability.

## Appendix 2. General Recipe: Developing a decision tree for model selection using database characteristics

Although we perform our analysis for a specific data/modeling context, the same basic "recipe" developed in this paper can be applied to many other settings. We formalize this recipe as a general method for model evaluation and selection, involving the three basic ingredients: the set of candidate models, the database characteristics, and the performance criterion. This enables the analyst to answer the questions "Which model should I use for this this dataset?" and "Given a dataset, how well will a given model perform?"

1. Models. Our procedure supposes that the analyst has a consideration set of models $1,\ldots,M$. The models can all be run on datasets with the same structure. We also suppose that model-based simulation can be done via Monte Carlo, or Markov Chain Monte

Carlo, if needed. Our procedure assumes that an MCMC sampler has been run to obtain draws $g = 1, \ldots, G$ from each model's the joint posterior distribution, $\left[\mathbf{\Omega}_m | \mathbf{Y}^{\text{obs}}\right]$.

2. Database characteristics. We characterize the database with a set of summary statistics. These should be (1) easy-to-compute characteristics, (2) managerially relevant, and (3) largely comprehensive and mutually exclusive. Formally, we denote these the dataset-level summary statistics as a covariate vector, $\tau(\mathbf{Y}_k^{\text{obs}})$, for dataset $k$. These are to be computed before running any models on the $k$th dataset, which itself is denoted by $\mathbf{Y}_k^{\text{obs}}$. These are the independent variables of interest in the eventual classification.

3. Performance criterion. Assessing model performance for model selection is an important step that should be driven by the business goal. We use an empirical validation approach via posterior predictive distributions. We generate data, $\mathbf{Y}_m^{*(g)}$, from the model-based predictive distribution, $\left[\mathbf{Y}_m^* | \mathbf{Y}^{\text{obs}}\right]$, where $g$ indexes replications $1, ..., G$.

   For the performance measure feature, $s$, we summarise the generated data by, $T_s(\mathbf{Y}_m^{*(g)})$. We quantify performance as model errors: the degree to which the model-based posterior predictive distribution of feature $s$ is outlying with respect to the feature's observed value, $T_s(\mathbf{Y}_k^{\text{obs}})$.

   Let $D$ denote a loss function along a single dimension, that is, the distance between the draws from the posterior predictive distribution and the single observed value of feature of a dataset. This distance, $d_{mks}$, summarizes model $m$'s performance on dataset $k$ in terms of feature $s$, utilizing all replicates $g = 1, \ldots, G$,

$$d_{mks} = D\left(T_s(\mathbf{Y}_{mk}^{*(1)}), \ldots, T_s(\mathbf{Y}_{mk}^{*(G)}); T_s(\mathbf{Y}_k^{\text{obs}})\right). \tag{15}$$

   The choice of the function $D$ should depend on the desired feature.

   Regardless of which performance metric and error measure is chosen, a single metric is obtained for each posterior replicate. For each replicate, we select the model with the lowest error and consider it the "winning" model, which is the nominal categorical outcome variable to be classified.

4. Classifying datasets by relating model performance to observed database characteristics. Putting those three ingredients together, we create the decision tree to infer the relationship between which model is best (outcome) and database characteristics (predictors). We formalize the classification as its own predictive tool. The independent variables of interest are the dataset-level summary statistics, $\tau(\mathbf{Y}_k^{\text{obs}})$, computed before running any models on the $k$th dataset, denoted by $\mathbf{Y}_k^{\text{obs}}$. These values are the predictors of model performance. The error measure, $d_{ksm}^{(g)}$, as defined above, is on a continuous scale. For classification purposes, however, the dependent variable should be an indicator of the "winning" model $m_{ks}^{\text{Winner (g)}}$, a nominal categorical variable with $M$ levels. For each dataset $k$, feature $s$, and posterior replicate $g$,

$$m_{ks}^{\text{Winner (g)}} = \text{argmin}_{m \in (1,\ldots,M)} \left\{ d_{ks1}^{(g)}, ..., d_{ksM}^{(g)} \right\}. \tag{16}$$

   We use a classification tree to relate the identity of the winning model, $m_k^{\text{Winner}}$, to the database characteristics, $\tau(\mathbf{Y}_k^{\text{obs}})$. Given the performance of all $M$ models across all $K$ datasets for feature $s$, we explain variations in the model performance (i.e., which model

wins) as a function of the observed summaries of that dataset. Formally, we capture this relationship as follows,

$$\hat{m}_k^{\text{Winner}(g)} = \text{Tree}\left[\tau(\mathbf{Y}_k^{\text{obs}})\right], \tag{17}$$

where "Tree" denotes the classification tree predicting the winning model $\hat{m}_k^{\text{Winner}(g)}$ for each of the datasets $k = 1, \ldots, K$ and replicates $g = 1, \ldots, G$. The results will show which dataset-level summaries are associated with differences in performance across the models for the feature of interest. The exact same setup used for CART methods can be used for implementing random forests. The same basic relationships are uncovered, but different methods are used.