COHORT REVENUE & RETENTION ANALYSIS: A BAYESIAN APPROACH

JUAN CAMILO ORDUZ

Abstract. We present a Bayesian framework for jointly modeling cohort-level retention and revenue over time. We contribute a framework that couples these two business metrics through the number of active users. We model retention non-parametrically using Bayesian additive regression trees (BART) and Bayesian neural networks to capture non-linear patterns and seasonality, and couple this with a gamma-distributed revenue model where the estimated active user counts from the retention component inform the revenue predictions. This approach enables information sharing across cohorts, naturally incorporates seasonal effects, and provides well-calibrated uncertainty quantification through highest density intervals. The framework is flexible enough to incorporate additional covariates (which can vary over time) in both model components. We demonstrate the approach using synthetic data designed to reflect realistic business scenarios, showing accurate out-of-sample predictions with appropriate uncertainty estimates. The framework's modular design facilitates extensions to hierarchical structures: we demonstrate a multi-market implementation that pools information across markets with varying data availability, enabling reliable forecasts even for markets with very limited cohort histories. Implementation code is provided in PyMC and NumPyro, where we use both MCMC and stochastic variational inference to fit the models, making the method accessible for practical applications.

1. Introduction

Understanding and predicting customer behavior directly impacts business profitability through improved retention strategies and resource allocation. Among the metrics that define business success, retention and customer lifetime value estimation stand at the forefront, serving as critical indicators of a company's ability to not only attract but maintain a loyal customer base. These metrics transcend mere financial accounting—they represent the foundation upon which long-term business strategies are built and refined. Seminal work by Fader and Hardie has established frameworks for both contractual settings [Fader and Hardie, 2007a], where subscription-based relationships predominate, and non-contractual settings [Fader et al., 2005], where customers may come and go without formal notification[1]. Modern implementations of these CLV models can now be found in Bayesian probabilistic programming frameworks such as PyMC ([Abril-Pla et al., 2023]). Specifically, the PyMC-Marketing library [PyMC-Labs, 2023] provides implementations of many standard buy-till-you-die (BTYD) models including the BG/NBD, Pareto/NBD, and Gamma-Gamma models in a flexible, Bayesian framework. While these approaches have proven very valuable, they often struggle to scale

Date: October 25, 2025.

[1]Our definition of retention corresponds to what they call survival curve. See precise definitions below.

effectively. They can definitively be scaled with modern hardware and algorithms (for example, stochastic variational inference, as described below). Nevertheless, this requires non-trivial work and effort.

For many decision-making processes, companies and senior management (think of a C-level executive) mostly need to understand behaviors at the cohort level—groups (e.g. cohorts of customers who joined during the same time period). In this paper we focus on this level of granularity. When shifting from individual to cohort-level analysis, businesses typically face a methodological trilemma:

(1) **Complete pooling**: Aggregate all cohorts together and model retention and revenue as a collective whole, potentially obscuring important cohort-specific patterns.
(2) **No pooling**: Analyze each cohort in isolation, potentially overlooking valuable cross-cohort information and suffering from data sparsity for newer cohorts.
(3) **Partial pooling**: Model cohorts jointly with shared parameters, striking a balance between cohort-specific insights and statistical power.

As detailed by [Fader and Hardie, 2017], each approach offers distinct advantages and limitations. However, a fundamental challenge persists across these traditional methodologies: they typically lack the flexibility to efficiently incorporate seasonality patterns and external regressors[2]. This limitation becomes particularly problematic for businesses with highly seasonal customer behavior—from retail operations affected by holiday shopping patterns to subscription services influenced by annual promotional cycles. While some might argue that seasonality is secondary when estimating customer lifetime value, the reality for many business models is that seasonal fluctuations significantly impact customer acquisition, engagement, and retention decisions. Beyond the methodological challenges, businesses face practical hurdles in translating retention and revenue models into actionable insights. Static models that fail to adapt to changing market dynamics or consumer preferences quickly become outdated. Moreover, point estimates without associated uncertainty measures can lead to misplaced confidence in business forecasts, potentially resulting in suboptimal resource allocation and strategic planning. The Bayesian cohort-revenue-retention framework presented in this paper addresses these challenges and has been successfully applied to real-world business datasets (both in contractual and non-contractual settings). As we will describe below, the model structure is flexible enough to incorporate extensive business specific prior knowledge through priors and convenient parametrizations. Most importantly, the framework provides a straightforward way to add custom covariates in both the retention and revenue components (and even in the coupling mechanism) which can vary over time. This is a fundamental advantage of this work, as in other classical probabilistic models like BG/NBD and Pareto/NBD, covariates can be added but are computationally expensive (and actually, the time-varying covariates require significant amoutn of work to implement in a way that scales).

[2]Although, one can add regressors in some cases as described in [Fader and Hardie, 2007b] for the non-contractual case.

Why Cohort-Level Modeling? Before proceeding further, let's come back to the beginning and address an important question: why focus on cohort-level rather than individual-level modeling? While individual-level models can provide granular predictions for specific customers, cohort-level analysis offers several strategic advantages that make it particularly well-suited for many business decision-making contexts.

- First, cohort-level aggregation substantially reduces noise inherent in individual transaction data. Individual purchase patterns are often highly variable and influenced by idiosyncratic factors that average out when aggregating to the cohort level. This noise reduction leads to more stable parameter estimates and more reliable forecasts, particularly valuable for strategic planning where robustness is paramount.
- Second, cohort-level models align naturally with how many business decisions are actually made. Marketing strategies, budget allocations, and resource planning typically target customer segments rather than individuals. Financial forecasting and business planning operate at aggregate levels. A model that directly addresses these cohort-level questions provides more immediately actionable insights than individual-level predictions that must subsequently be aggregated. This alignment enhances accessibility for marketing stakeholders and strategic decision-makers who may not require customer-specific granularity but need to understand temporal patterns and cohort dynamics. For example, the model presented in this paper sucessfully enabeled data-driven decisions at a company operating in multiple countries where user-level predictions were hard to understand and make actionable. It was strategically better to segment the cohorts (adding other dimensions like marketing aquisition channel) to extract relevant signals and insights for senior management.
- Third, from a computational and practical perspective, cohort-level models scale more favorably than individual-level approaches, particularly as customer bases grow into millions. While modern Bayesian methods can handle large individual-level datasets, the computational requirements and implementation complexity increase substantially. Cohort-level modeling offers an efficient path to actionable insights without sacrificing the flexibility to incorporate rich covariate information when needed.

Importantly, the cohort-level focus does not preclude the incorporation of customer characteristics. As we demonstrate in our framework, additional covariates—such as acquisition channel, geographic location, or customer segment—can be seamlessly integrated into our model. To capture complex patterns in retention behavior without requiring explicit specification of functional forms, we employ Bayesian Additive Regression Trees (BART) [Chipman et al., 2010]. BART is a flexible non-parametric method that represents the unknown function as a sum of regression trees, allowing it to automatically learn non-linear relationships and interactions between features. This flexibility is particularly valuable for cohort retention modeling, where relationships between temporal features (cohort age, calendar time, seasonality) may be complex and difficult to specify a priori. The non-parametric nature of BART allows it to scale effectively with

many features, enabling the model to capture heterogeneous effects across different customer types while maintaining the interpretability advantages of cohort-level aggregation through tools like partial dependence plots. Thus, our approach strikes a balance: operating at the cohort level for strategic clarity while retaining the flexibility to incorporate individual-level characteristics when they provide additional explanatory power. Moreover, as we can additional imporse a hierarchical structure across markets or subset of covariates, we can efficiently pool information across different cohorts and markets. Trying to do this pooling at the individual level is essentially impossible to a scale of millions of customers (which is more the norm than the exception).

Related Work and Literature. Our approach sits at the intersection of several research streams in customer analytics, survival analysis, and cohort modeling. Understanding how our contribution relates to and extends existing methodologies is crucial for appreciating its novelty and practical value.

Age-Period-Cohort Models. The statistical literature on age-period-cohort (APC) modeling provides a rich framework for understanding temporal patterns in grouped data. As comprehensively reviewed by [Fannon and Nielsen, 2018], APC models decompose outcomes into effects attributable to age (time since an event), period (calendar time effects), and cohort (group membership defined by a common temporal characteristic). These models have found extensive application in demography, epidemiology, and social sciences. However, traditional APC approaches face the well-known identification problem: age, period, and cohort are linearly dependent (cohort + age = period), making their individual effects non-identifiable without additional constraints. Moreover, standard APC models are typically specified for a single outcome variable, with additive decomposition of age, period, and cohort effects. While APC models have been extended to multivariate settings in some contexts, the standard framework and most applications focus on univariate outcomes. Our framework shares with APC models the recognition that outcomes depend on age (time since cohort formation), period (calendar time), and cohort identity. However, rather than focusing on decomposing and identifying separate age, period, and cohort effects—which requires imposing constraints or noting that only non-linear components are identifiable—we use flexible non-parametric modeling with BART that captures the joint functional relationship between these temporal dimensions without requiring explicit decomposition. Furthermore, we extend to joint modeling of two related outcomes (retention and revenue) through a principled coupling mechanism.

Survival Analysis and Retention Modeling. In the survival analysis literature, [Hubbard et al., 2021] recently introduced Beta Survival Models that use non-parametric methods (including tree-based approaches) to model discrete-time survival probabilities with a beta-logistic formulation. Their work demonstrates the value of flexible, non-parametric approaches for capturing heterogeneous survival patterns and forecasting beyond observed horizons—capabilities that are particularly relevant for retention modeling in business contexts. Our retention component shares the motivation of using flexible non-parametric methods (BART in our case) to model time-to-churn patterns. However, we extend this foundation in two critical directions. First, while [Hubbard et al., 2021] focus solely on survival/retention, we introduce a novel coupling mechanism that connects retention to

revenue through the number of active users, enabling joint forecasting of both business-critical metrics. Second, our framework explicitly incorporates cohort structure and temporal effects in a way that facilitates information sharing across cohorts—a feature absent from standard survival models but essential for business applications where newer cohorts have limited historical data.

Customer Lifetime Value Models. The customer lifetime value literature, pioneered by the work of Fader and Hardie on buy-till-you-die (BTYD) models [Fader et al., 2005, Fader and Hardie, 2007a], provides powerful frameworks for individual-level customer behavior modeling. These approaches, particularly the BG/NBD and Pareto/NBD models, have become standard tools in marketing analytics. Modern Bayesian implementations in packages like PyMC-Marketing [PyMC-Labs, 2023] have made these models more accessible and extended their capabilities. However, as previously noted, BTYD models operate primarily at the individual level and can face scalability challenges. While [Fader and Hardie, 2007b] demonstrates that covariates can be incorporated, and [Fader and Hardie, 2017] discusses multi-cohort fitting strategies, these extensions still work within the constraints of the original parametric model structures. Our approach complements this literature by offering an alternative perspective: rather than aggregating individual-level predictions, we directly model cohort-level patterns using flexible non-parametric methods that naturally capture complex interactions between temporal effects, seasonality, and cohort characteristics.

Research Gaps and Our Contributions. The existing literature reveals several gaps that our work addresses. First, while age-period-cohort models provide a framework for temporal decomposition, they lack the flexibility to capture non-linear patterns and are typically not designed for joint modeling of multiple related outcomes. Second, survival analysis approaches focus on single outcomes (survival/retention) and do not provide mechanisms for connecting retention to downstream business metrics like revenue. Third, traditional CLV models, while powerful for individual-level prediction, can struggle with computational scalability and may not naturally incorporate rich temporal patterns like seasonality without significant modeling effort. Our framework addresses these gaps through: (1) flexible non-parametric modeling via BART that captures complex temporal patterns without requiring explicit functional form specification, (2) a novel coupling mechanism that jointly models retention and revenue while preserving interpretability, (3) principled Bayesian uncertainty quantification through posterior distributions rather than point estimates, and (4) efficient cohort-level aggregation that balances statistical power with practical scalability. In the following sections, we formalize this framework and demonstrate its effectiveness.

Having established the motivation and positioned our work within the broader literature, we now turn to describing the modeling approach in detail. To get a visual intuition of the data we want to model, Figure 1 shows an example of a retention matrix. Here we encode the cohort retention as a function of time. Note that we exclude the diagonal as it is uninformative (always containing ones). Observe that older cohorts have more data (obviously), so we would like to use this information to improve the estimation of retention for younger cohorts. Hence, we do not want to model each cohort independently but rather the *whole retention matrix* (we will do the same for the revenue matrix and

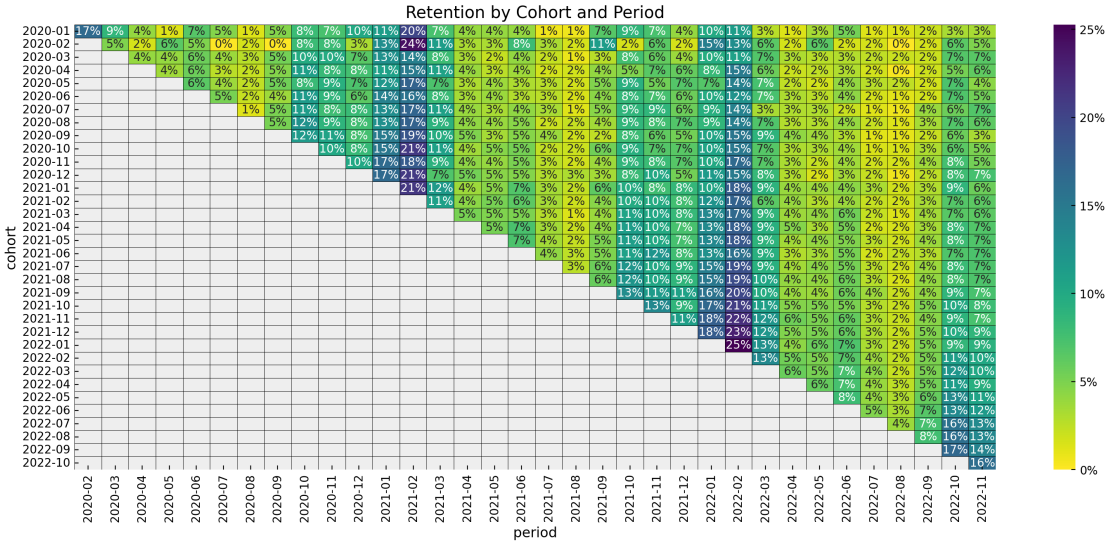

FIGURE 1. Retention matrix example. The matrix visualizes customer retention rates across different cohorts (rows) and observation periods (columns). Each cell represents the proportion of customers from a specific acquisition cohort that remained active in a subsequent period. Colors indicate retention rates, with darker colors typically showing higher retention. This visualization allows for identifying cohort-specific patterns, seasonal effects, and retention decay over time. The diagonal is excluded as it always contains trivial values of 1 (100% retention) for the cohort's first period.

couple them together).

In addition, as we want to understand the monetary contribution of each cohort, we can consider the revenue matrix as shown in Figure 2. As in the retention case, we want to make sure we use all the information available to improve the estimation of revenue for younger cohorts. Moreover, as we will discuss below, we will couple the retention and revenue matrices through the number of active users, making the model structure very transparent for the business users and stakeholders.

This approach offers several distinct advantages:

- **Flexibility in relationship modeling**: By employing Bayesian additive regression trees (BART) [Quiroga et al., 2022], our approach can capture complex non-linear relationships between cohorts, time periods, and behavioral metrics without requiring explicit specification of these relationships.
- **Integrated seasonality**: The model naturally incorporates seasonal patterns without requiring separate components or preprocessing steps.
- **Extensibility**: Additional covariates—from macroeconomic indicators to marketing campaign intensities—can be seamlessly integrated into the model.

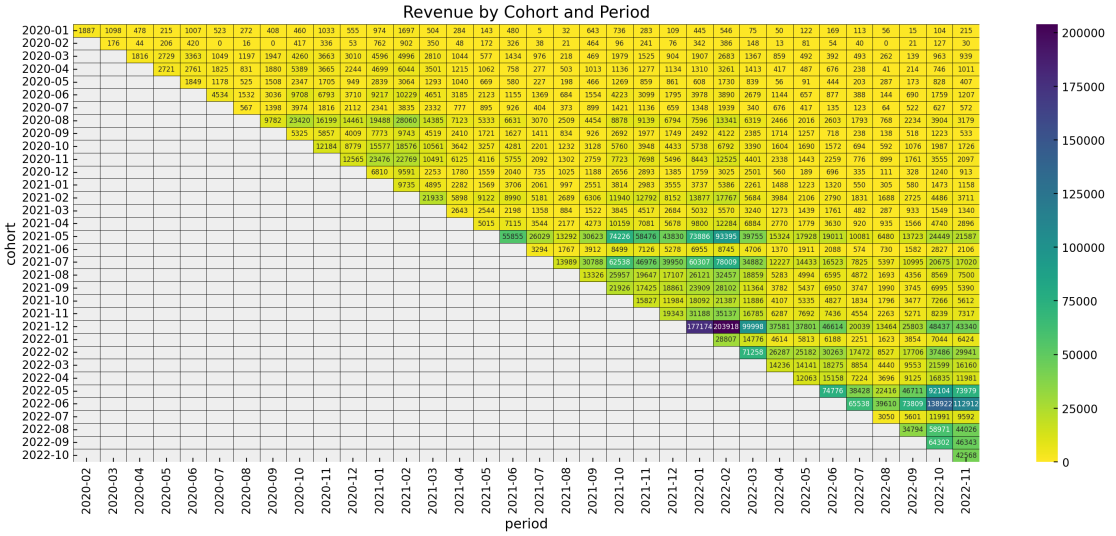

FIGURE 2. Revenue per cohort. This heatmap visualizes the total revenue generated by each cohort (rows) across different time periods (columns). The color intensity corresponds to revenue magnitude, revealing a strong correlation with the number of active users (Figure 3).

- **Uncertainty quantification**: The Bayesian framework provides natural uncertainty estimates around all predictions, enabling risk-aware decision making.
- **Information sharing across cohorts**: Newer cohorts with limited historical data benefit from patterns learned from more established cohorts.

An important strength of this modular framework is its extensibility to more complex structures. To illustrate, we present a hierarchical extension that models multiple markets simultaneously, pooling information across them. This is particularly valuable when some markets have limited data: by sharing information through hierarchical priors, we can generate reliable forecasts for young markets that would be impossible to model in isolation. The hierarchical implementation, detailed in Section 6, demonstrates how naturally the framework accommodates such extensions without fundamental redesign.

Specifically, we use Bayesian additive regression trees to model the retention component, capturing the probability that a customer from a given cohort remains active in subsequent periods. We couple this with a linear model for the revenue component, predicting how much revenue active customers will generate. This dual approach balances the flexibility needed to capture complex retention patterns with the interpretability desired for revenue forecasting. Next we describe the main ingredients of our model: the features and the model specification. We will delve into the details in the next sections.

Features. Typical purchase databases contain transactional history at user level. We want an approach general enough to benefit from the most common features instead of heavy feature engineering. Going back to Figure 1, it is natural to consider the following features to model the retention and revenue matrices:

- **Cohort age**: Age of the cohort in months, representing the time since the cohort was formed.
- **Age**: Age of the cohort with respect to the observation time. This feature serves as a numerical encoder for the cohort's position in time.
- **Month**: Month of the observation time (period), capturing seasonality effects.

For example, if our observation month is *2022-11* and we consider the cohort *2022-09*, the age of this cohort is 2 months, as the age is always calculated relative to the observation period. This cohort was observed during two periods: *2022-10* and *2022-11* with cohort ages 1 and 2 respectively.

All these features are available for out-of-sample predictions, ensuring model applicability for forecasting. In practice, we can add additional covariates to the model. The only requirement for out-of-sample predictions is that these covariates must be available for future observation periods.

Model Specification. The main idea behind the specification is to model each revenue and retention matrices, using the features above, and couple them together. Specifically, we have:

- **Retention Component**: We model the number of active users N_{active} in each cohort as a binomial random variable $\text{Binomial}(N_{\text{total}}, p)$, where the parameter p represents the retention probability (see Figure 3, from a synthetic example described below). We model the latent variable p using a BART model with features cohort age, age, and month. This flexible approach allows the model to capture non-linear relationships and interactions between features.
- **Revenue Component**: We model the revenue matrix (see Figure 2) through a gamma random variable $\text{Gamma}(N_{\text{active}}, \lambda)$, as we want to ensure non-negative values. We model the rate parameter λ through a linear model with features cohort age, age, and a multiplicative interaction term (using a log link function). We do not explicitly add a seasonality component to this part of the model, as we typically observe that most seasonality effects are already captured by the retention component. However, seasonal features could be added if needed (plus additional features and different parametrizations, for example multiplicative effects).
- **Coupling**: The retention and revenue coupling is a key feature of this work. We couple the two components through the number of active users, extending decomposition ideas similar to those in the Gamma-Gamma model from the CLV literature, which separates transaction frequency from monetary value. Here is the full model specification:

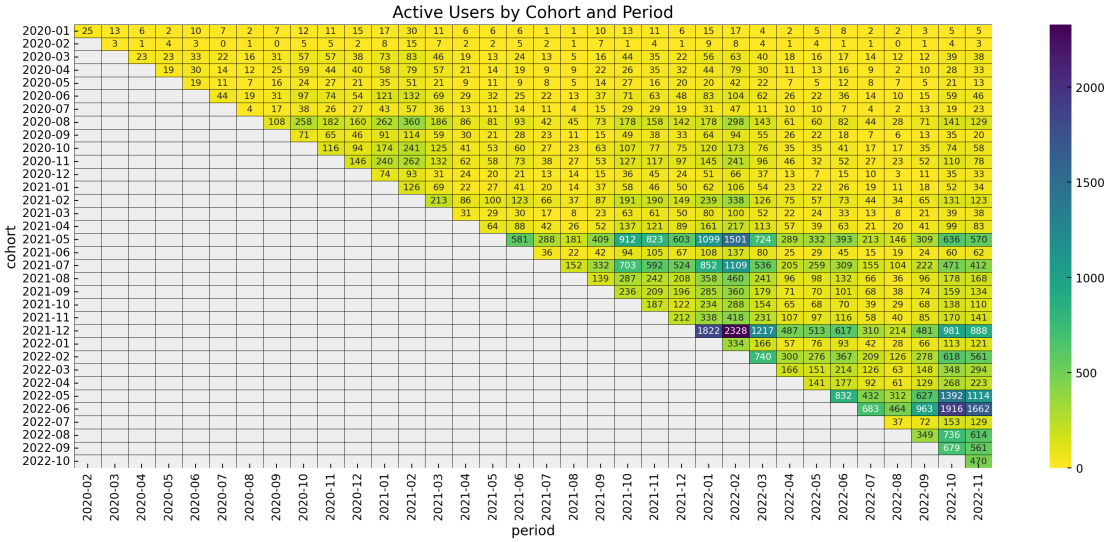

FIGURE 3. Number of active users across cohorts. This heatmap displays the absolute count of active users for each cohort (rows) across observation periods (columns).

$$
\begin{aligned}
\text{Revenue} &\sim \text{Gamma}(N_{\text{active}}, \lambda) \\
\log(\lambda) &= (\text{intercept} \\
&\quad + \beta_{\text{cohort age}} \times \text{cohort age} \\
&\quad + \beta_{\text{age}} \times \text{age} \\
&\quad + \beta_{\text{cohort age} \times \text{age}} \times \text{cohort age} \times \text{age}) \\
N_{\text{active}} &\sim \text{Binomial}(N_{\text{total}}, p) \\
\text{logit}(p) &= \text{BART}(\text{cohort age}, \text{age}, \text{month})
\end{aligned}
$$

Figure 4 illustrates the complete model structure. Our goal is to simultaneously estimate the BART parameters and the beta coefficients (including the intercept) of the linear component. We want to do this to understand the contribution of each feature to the retention and revenue over time. Additionally, to operationalize the model, we will use the retention and revenue matrices to make out-of-sample predictions. This can be extremely important for scenario and business planning. A typical application is to use this model to generate *counterfactuals* for global interventions where we expect different cohorts to react differently.

In the rest of the paper, we delve into the details of the model specification and diagnostics. Moreover, we describe how to generate out-of-sample predictions for both the retention and revenue matrices.

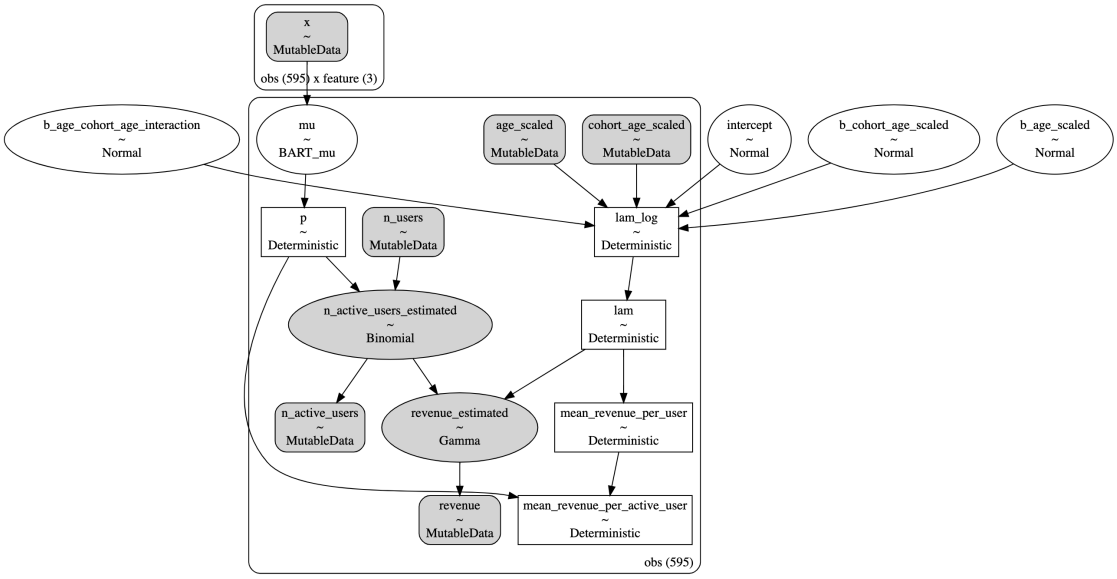

FIGURE 4. Cohort-revenue-retention model structure. This diagram illustrates the coupling mechanism that connects the two components of our framework. On the left, the retention component models the number of active users N_{active} as a binomial random variable, where the retention probability p is modeled using BART with features including cohort age, age (cohort identifier), and month (seasonality). On the right, the revenue component models total revenue as a gamma-distributed random variable, with the shape parameter directly determined by N_{active} from the retention model. This coupling through active users provides a natural connection: changes in retention patterns automatically propagate to revenue predictions, ensuring consistency between the two metrics while allowing each component to use appropriate distributional assumptions and feature sets.

To make the approach more tangible, we present a synthetic dataset in the next section. This should help the reader to better understand the data and the approach.

2. Synthetic Data

Having established the theoretical foundations and positioning of our approach, we now turn to a concrete demonstration using synthetic data. While our framework has been successfully applied to real business datasets, we present results using synthetic data for reproducibility and to avoid proprietary information concerns. The synthetic dataset is designed to reflect realistic business scenarios, incorporating the types of temporal patterns and cohort dynamics commonly observed in subscription-based and retention-focused business models.

The synthetic dataset is available as a CSV file from [Orduz, 2023b], and the code to generate this dataset deterministically is publicly available in [Orduz, 2023c]. This

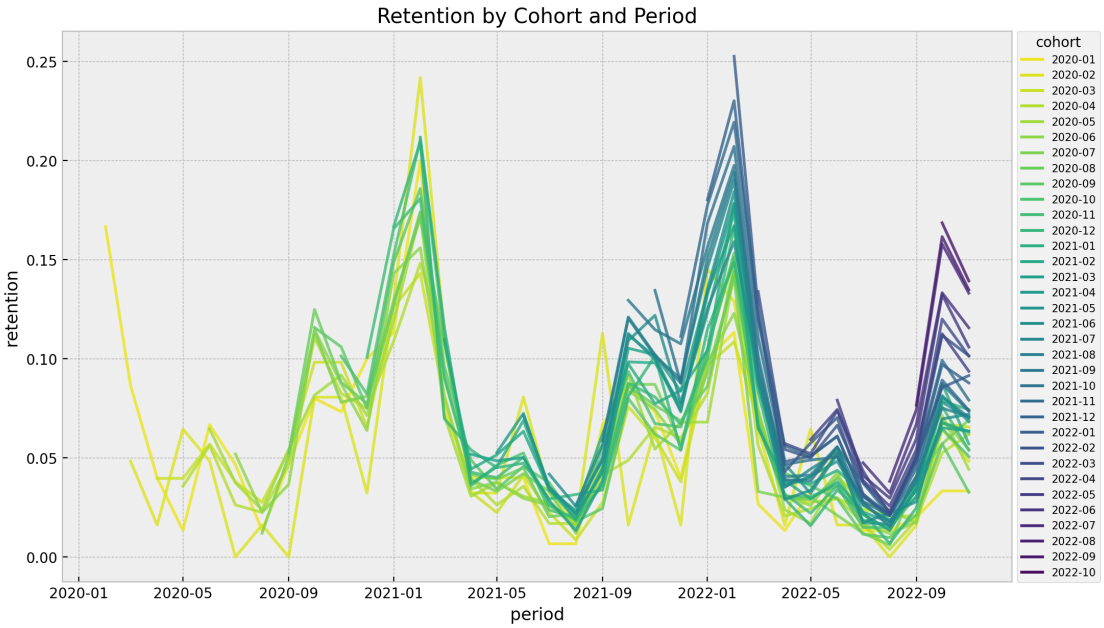

FIGURE 5. Retention as a function of the period, demonstrating the yearly seasonality pattern in retention values.

ensures full reproducibility of our results and allows researchers to explore model behavior under controlled conditions.

Exploratory Data Analysis. Before fitting our model, we conduct exploratory data analysis to understand the key patterns in the data. This analysis both motivates our modeling choices and provides a baseline for evaluating model performance. Figure 1 displays the retention matrix per cohort and period. Two key observations stand out:

(1) The retention exhibits a clear seasonal pattern with respect to the period, being higher in the last months of the year and lower in the middle of the year. This seasonality pattern is more evident in Figure 5.
(2) Retention appears to increase as the cohort age decreases. This trend is apparent when comparing retention values for periods in November across different cohort ages.

It's important to remember that retention is a ratio, making cohort size an important factor. For instance, a retention rate of 0.4 could represent either $4/10$ or $4 \times 10^5/10^6$. The former case carries considerably more uncertainty in its estimation. This insight motivates us to examine the number of active users, as shown in Figure 3. We observe that more recent cohorts have significantly more active users, a pattern we want our model to account for.

Next, we examine revenue patterns. Figure 2 presents revenue by cohort, showing a strong correlation with the number of active users. This suggests that revenue per user

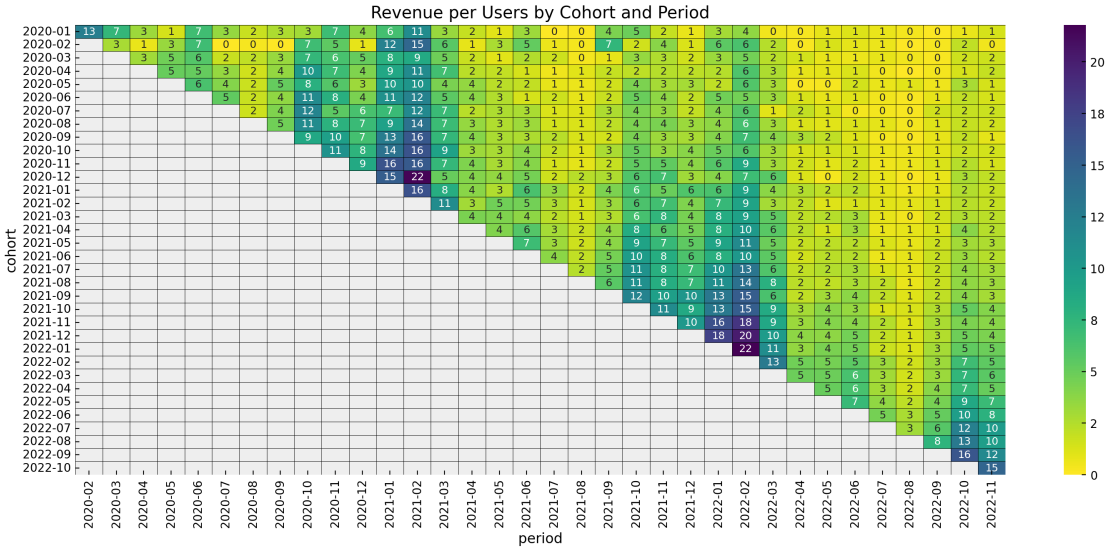

FIGURE 6. Revenue per user across cohorts. This visualization normalizes the total revenue by the original cohort size, showing the average revenue generated per initially acquired user.

remains relatively stable over time. To verify this, we compute revenue per user as a function of age and period (Figure 6) as well as revenue per *active* user (Figure 7). The key difference between these metrics is that revenue per user divides by total cohort size, while revenue per active user divides by the number of active users in the given period. All in all, we observe the following for the revenue data[3]:

- Revenue per user exhibits a clear seasonality pattern, consistent with the seasonal pattern observed in retention.
- Revenue per active user does not show the same seasonality pattern since seasonal effects are already captured in the denominator (active users). Additionally, revenue per active user appears to decrease as cohort age increases, suggesting that older cohorts generate less revenue per active customer.

These exploratory findings inform our modeling strategy. The strong seasonality in retention motivates the inclusion of month (period) as a feature in the BART component. The heterogeneity across cohort ages suggests that flexible non-parametric modeling will be valuable for capturing these varying patterns. The relationship between revenue and active users justifies our coupling mechanism, while the patterns in revenue per active user guide the parametric specification of the revenue component. With this exploratory understanding established, we now proceed to formalizing and fitting our model.

[3]These types of patterns are actually common in real applications. This synthetic dataset is motivated by real applications where the model was proven to be very effective

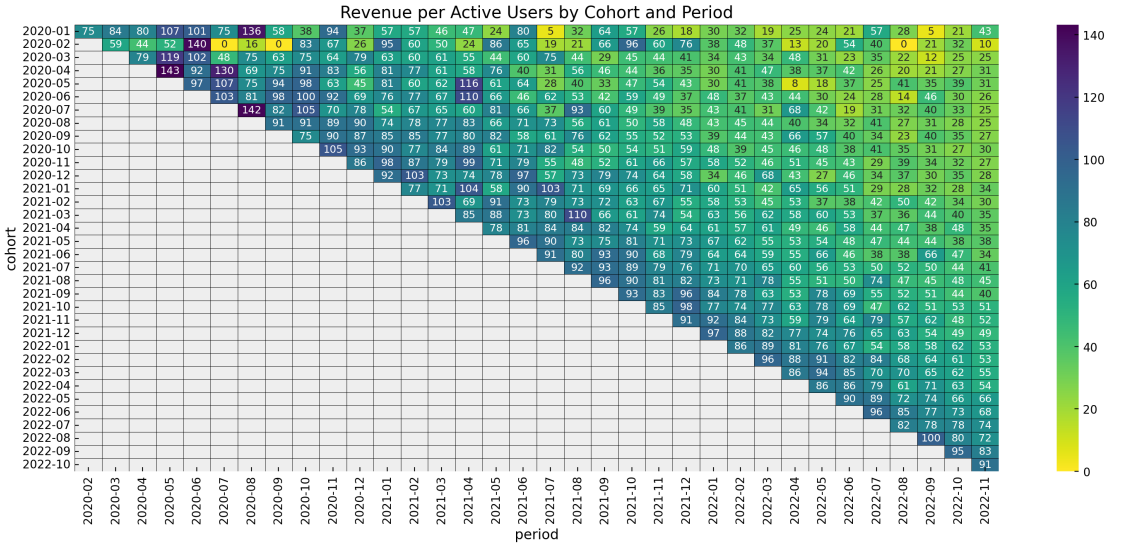

FIGURE 7. Revenue per active user across cohorts. This metric divides total revenue by the number of active users in each period, isolating spending patterns from retention effects.

3. Model Structure and Diagnostics

3.1. **Model Specification.** Let's expand on the model structure outlined in the introduction. The core concept is to model the number of active users as a binomial random variable $N_{\text{active}} \sim \text{Binomial}(N_{\text{total}}, p)$, where p represents the retention probability. We use Bayesian additive regression trees (BART) to model this latent variable p using cohort age, age, and month (period) as features.

$$N_{\text{active}} \sim \text{Binomial}(N_{\text{total}}, p)$$
$$\text{logit}(p) = \text{BART}(\text{cohort age}, \text{age}, \text{month})$$

It is important to clarify the role of N_{active} in the model specification versus estimation. In the model formulation above, N_{active} is formally treated as a random variable following a binomial distribution. However, in practical applications with observed cohort data, we directly observe retention rates (the proportion of customers who remain active) or equivalently, the realized counts N_{active}. During inference, the observed retention data informs the estimation of p through the BART model, which captures how retention probability varies across cohorts, time periods, and months. This formulation allows us to propagate uncertainty from the retention probability estimates through to revenue predictions via the coupling mechanism described below.

BART Prior Specification. BART models the unknown retention function as a sum of many regression trees, where each tree contributes a small part to the overall prediction [Chipman et al., 2010]. The key advantage is that this ensemble automatically learns complex non-linear patterns and interactions without requiring the analyst to specify

functional forms. The prior specification encourages sparse, shallow trees that collectively capture the data structure while avoiding overfitting.

We implement BART using the PyMC framework via [Quiroga et al., 2022], which handles the technical details of tree structure priors and parameter estimation. The main tuning parameter is the number of trees m. We use default prior settings from [Chipman et al., 2010] which have proven robust across diverse applications: these priors favor shallow trees and regularize individual tree contributions so that the ensemble learns smooth functions through additive combinations. For our retention modeling, we typically start with $m = 20$ trees and increase incrementally (e.g., to 50 or 100) while monitoring posterior predictive fit. BART is relatively insensitive to this choice for sufficiently large m, but smaller values offer computational efficiency and maintain interpretability through tools like partial dependence plots.

Readers interested in the mathematical details of BART priors—including tree topology priors, splitting rule distributions, and leaf parameter specifications—are referred to the comprehensive treatment in [Chipman et al., 2010] and the implementation details in [Quiroga et al., 2022].

Remark 1. A key advantage of the BART model is its flexibility in incorporating additional covariates. In real business applications, we have successfully added customer segmentation features (such as acquisition media channels from attribution models). This provides valuable insights into media channel return-on-investment (ROI), allowing businesses to consider not just acquisition costs but also estimated customer lifetime value through this combined model.

Remark 2. While one could start with a simpler model, such as a linear model as described in [Orduz, 2023b], our experience with real datasets shows that such simpler approaches often fail to adequately capture the complex patterns in the data.

Computational Details and Inference. Inference for BART models is performed using Markov Chain Monte Carlo (MCMC) methods, specifically the Particle Gibbs sampler implemented in [Quiroga et al., 2022]. This algorithm iteratively updates tree structures through propose-accept-reject steps and updates leaf parameters through Gibbs sampling conditional on the tree structures. The MCMC sampler requires specification of the number of chains, number of draws, and warm-up (burn-in) period. We use multiple chains (typically 4) to assess convergence through standard diagnostics. For the synthetic data analysis presented in this paper, we run 2000 draws with 1000 warm-up iterations per chain, which proves sufficient for convergence. Larger datasets or more complex models may require more iterations.

Convergence diagnostics play a crucial role in ensuring reliable inference. We monitor the $\hat{R}$ statistic (Gelman-Rubin diagnostic) for all parameters, requiring $\hat{R} < 1.01$ for convergence. We also visually inspect trace plots for the linear model parameters to ensure proper mixing and stationarity. The BART component's stochastic nature (sampling tree structures) means individual tree parameters are not directly interpretable, but we can assess overall model convergence through the posterior predictive distribution and by monitoring whether different chains produce similar predictions. We verify that the

model produces no divergent transitions, which would indicate problematic posterior geometry. All diagnostics are implemented using the ArviZ library for exploratory analysis of Bayesian models.

For the revenue component, we employ a gamma random variable $\text{Gamma}(N_{\text{active}}, \lambda)$ (inspired by [Stucchio, 2015]). The gamma distribution is a natural choice for modeling revenue as it ensures non-negativity and provides flexibility in capturing different revenue distributions through its shape and rate parameters. The mean of this gamma distribution is $N_{\text{active}}/\lambda$, allowing us to interpret $1/\lambda$ as the *average revenue per active user*. By using N_{active} as the shape parameter, we ensure that cohorts with more active users have lower relative variance (coefficient of variation $1/\sqrt{N_{\text{active}}}$), which aligns with the intuition that aggregated revenue from larger cohorts should be more stable. This parametrization also creates a direct connection between the retention and revenue components: cohorts with higher retention (larger N_{active}) contribute both higher expected revenue and more concentrated revenue distributions. We model $\log(\lambda)$ using a linear function of cohort age, age, and their interaction. As a preprocessing step, we standardize these features for the linear model component (we keep the same notation for the variables for simplicity). This allows us to specify priors for the regression coefficients in terms of the effect of a one-standard-deviation change in the predictor, enabling effective regularization through standard normal priors for the coefficients (see [Orduz, 2023a]).

$$
\begin{aligned}
\text{Revenue} &\sim \text{Gamma}(N_{\text{active}}, \lambda) \\
\log(\lambda) &= (\text{intercept} \\
&\quad + \beta_{\text{cohort age}} \times \text{cohort age} \\
&\quad + \beta_{\text{age}} \times \text{age} \\
&\quad + \beta_{\text{cohort age}\times\text{age}} \times \text{cohort age} \times \text{age})
\end{aligned}
$$

A key insight from both this synthetic dataset and many real-world applications is that we typically don't need to explicitly model seasonality in the revenue component, as seasonal patterns are already captured by the retention component.

Remark 3. The *age* feature characterizes each cohort's temporal position. While we could replace this numerical encoding with a one-hot encoding of cohorts and add hierarchical structure to pool information across cohorts, the numerical encoding is more parsimonious under the assumption that temporally proximate cohorts behave more similarly than distant ones.

In summary, our cohort-revenue-retention model is specified as:

$$
\begin{aligned}
\text{Revenue} &\sim \text{Gamma}(N_{\text{active}}, \lambda) \\
\log(\lambda) &= (\text{intercept} \\
&\quad + \beta_{\text{cohort age}} \times \text{cohort age} \\
&\quad + \beta_{\text{age}} \times \text{age} \\
&\quad + \beta_{\text{cohort age}\times\text{age}} \times \text{cohort age} \times \text{age}) \\
N_{\text{active}} &\sim \text{Binomial}(N_{\text{total}}, p) \\
\text{logit}(p) &= \text{BART}(\text{cohort age}, \text{age}, \text{month}) \\
\text{intercept} &\sim \text{Normal}(0, 1) \\
\beta_{\text{cohort age}} &\sim \text{Normal}(0, 1) \\
\beta_{\text{age}} &\sim \text{Normal}(0, 1) \\
\beta_{\text{cohort age}\times\text{age}} &\sim \text{Normal}(0, 1)
\end{aligned}
$$

3.2. **Diagnostics.** With the model fully specified, we implement and fit it using PyMC [Abril-Pla et al., 2023], leveraging the BART implementation from [Quiroga et al., 2022]. Complete implementation details and code are available in [Orduz, 2023b]. After fitting the model using MCMC with the computational settings described above, we conduct thorough diagnostics to ensure reliable inference.

Figure 8 presents a critical diagnostic: the posterior predictive distribution for both model components. These plots compare the distribution of observed values against the distribution of values simulated from the fitted model's posterior predictive distribution. Close agreement between these distributions indicates that the model successfully captures the data-generating process. For both retention and revenue components, we observe excellent agreement, with the simulated distributions (orange) closely matching the observed distributions (black). This suggests the model provides an adequate fit to the data.

Beyond the posterior predictive check, we examine convergence diagnostics for the model parameters. Figure 9 displays trace plots for the linear model parameters (intercept and regression coefficients). These plots show the evolution of parameter values across MCMC iterations for each chain. Good mixing—evidenced by chains that explore the parameter space efficiently without getting stuck—is essential for reliable inference. We observe healthy mixing for all parameters, with no divergences or convergence warnings. All $\hat{R}$ statistics are below 1.01, confirming convergence. These diagnostics give us confidence that the MCMC sampling has successfully explored the posterior distribution and that our parameter estimates are reliable.

3.3. **Variable Importance.** A key advantage of using BART over more opaque machine learning methods is the availability of tools for understanding which features drive predictions and how they influence outcomes. The BART implementation in [Quiroga et al., 2022] provides interpretability tools that allow us to peer inside the model and extract actionable insights about retention drivers.

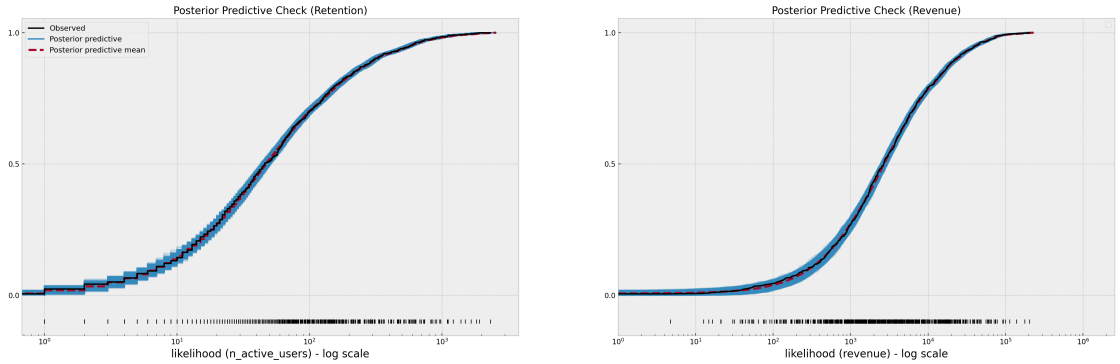

FIGURE 8. Posterior predictive distribution of the retention (left) and revenue (right) components, showing good fit to the observed data. These cumulative density plots compare the distributions of observed values (black) with simulated values from the posterior predictive distribution (orange), providing a visual assessment of model fit.

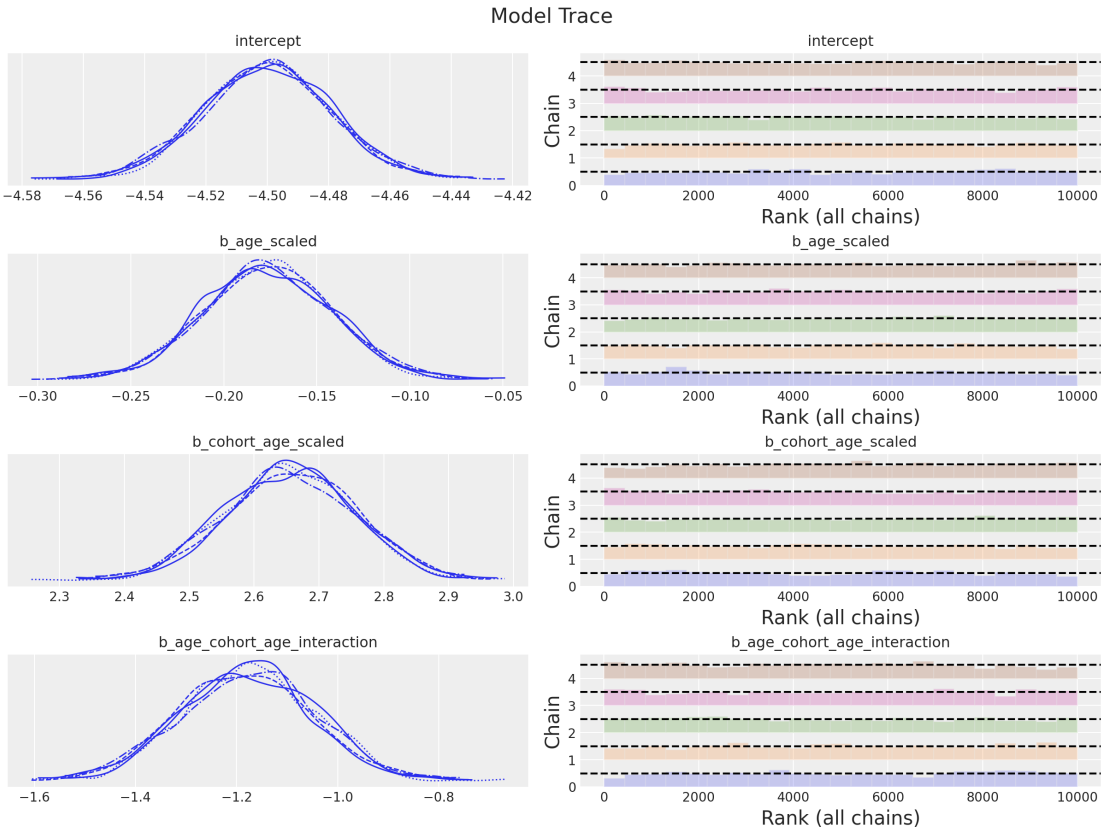

FIGURE 9. Trace plots for the linear model parameters, showing good mixing and convergence of the MCMC chains.

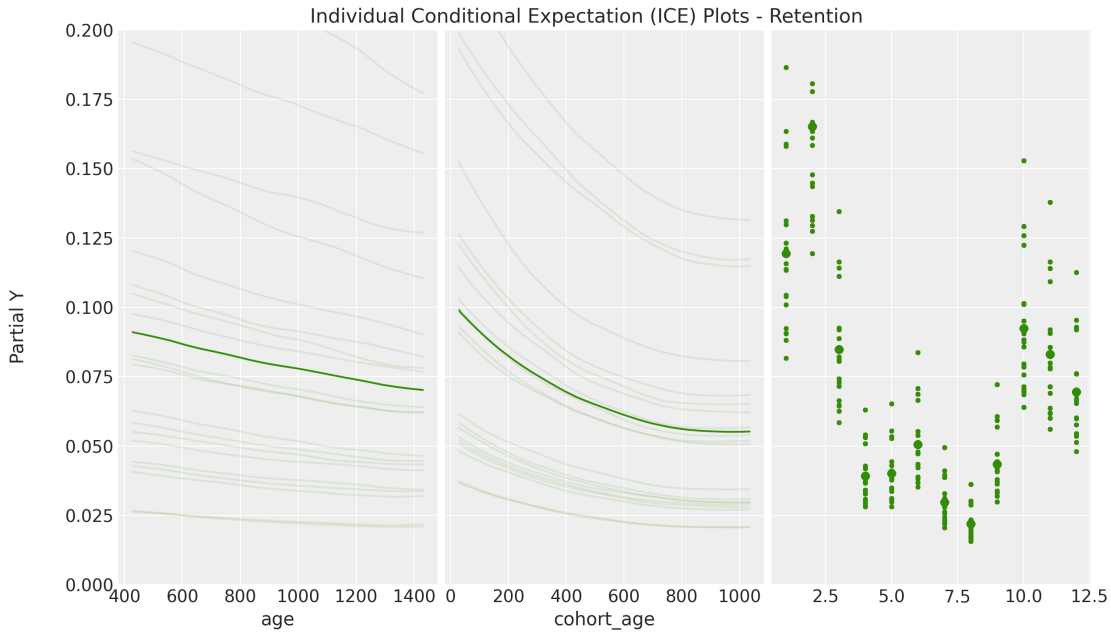

FIGURE 10. Partial Dependence Plot (PDP, solid lines) and Individual Conditional Expectation (ICE, thin dashed lines) plots for the retention component. These plots visualize how the predicted retention probability changes as each feature varies while holding others constant. The left panel shows the effect of cohort age: retention decreases as cohorts mature, with relatively consistent patterns across observations. The middle panel displays the age (cohort identifier) effect: more recent cohorts tend to have higher retention. The right panel reveals strong seasonality: retention peaks in months 11-12 (November-December) and dips in mid-year months. The close alignment between individual ICE plots and the average PDP suggests relatively consistent feature effects across different observations, indicating limited interaction effects. These plots demonstrate one of BART's key advantages: interpretable visualization of learned patterns without requiring pre-specification of functional forms.

Figure 10 presents PDP (Partial Dependence Plot) and ICE (Individual Conditional Expectation) plots for the retention component. These visualization techniques reveal how the model's predictions change as each feature varies while holding all other features constant. Each line represents a different observation from the dataset, showing how the predicted retention probability would change for that observation if we modified only the feature of interest. The PDP plot is the average of the ICE plots (solid line). These plots allow us to understand how the retention probability varies for different values of the features and reveals potential non-linear relationships or interaction effects that might not be apparent in aggregate statistics.

In this specific example, we can extract the following insights:

- The ICE plots show how the retention rate decreases with both cohort age and age. This is not surprising as we saw in the EDA.
- We see that the ICE plots have a similar trend to the PDP plots. This hints that the interaction effects are not so important in this case. This is also something we saw in the linear model where the interaction coefficient was relatively small (see [Orduz, 2022]).
- We clearly see the seasonality component of the PDP / ICE plots resemble the regression coefficients in the linear model from [Orduz, 2022]. This is simply representing the strong seasonal component of the data.

In addition, we can extract a relative importance for the different features using the contribution to the in-sample R^2, as shown in Figure 11.

These types of plots are very valuable to understand the *drivers* of the retention component.

4. Predictions

Having established that our model fits the data well and that MCMC sampling has converged, we now examine the model's predictive performance. This section evaluates predictions in two contexts: in-sample predictions that assess how well the model captures observed patterns, and out-of-sample predictions that test the model's ability to forecast future retention and revenue for periods beyond the training data. The distinction is crucial: in-sample fit demonstrates that the model adequately represents the data-generating process, while out-of-sample performance reveals whether the model generalizes to new data—the ultimate test for any predictive framework intended for business forecasting.

4.1. **In-Sample Predictions.**

We first evaluate the model's in-sample performance by comparing the posterior predictive mean against the observed values. Figure 12 shows the comparison for both retention and revenue components, with points closer to the diagonal line indicating better fit. Beyond point estimates, we can visualize the full posterior predictive distribution to assess model uncertainty. Figure 13 shows the posterior predictive distribution of retention for selected cohorts, with 94% HDI (Highest Density Interval). Note how the intervals are narrower for more recent cohorts with more data, reflecting greater certainty in these predictions. Overall, the predictions effectively capture the observed retention patterns, including seasonality. For the revenue component, Figure 14 shows the posterior predictive distribution compared to actual revenue values. The model successfully captures the revenue variability across different cohorts and time periods. We can use the whole posterior distribution to make custom visualizations of quantities of interests like the revenue per active user, as shown in Figure 15.

4.2. **Out-of-Sample Predictions.**

The true test of any predictive model is its performance on unseen data. We evaluate our model's forecasting capabilities using a holdout set consisting of data after *2022-11*, which was not used during model training. Figures 16 and 17 show the out-of-sample predictions for retention and revenue, respectively.

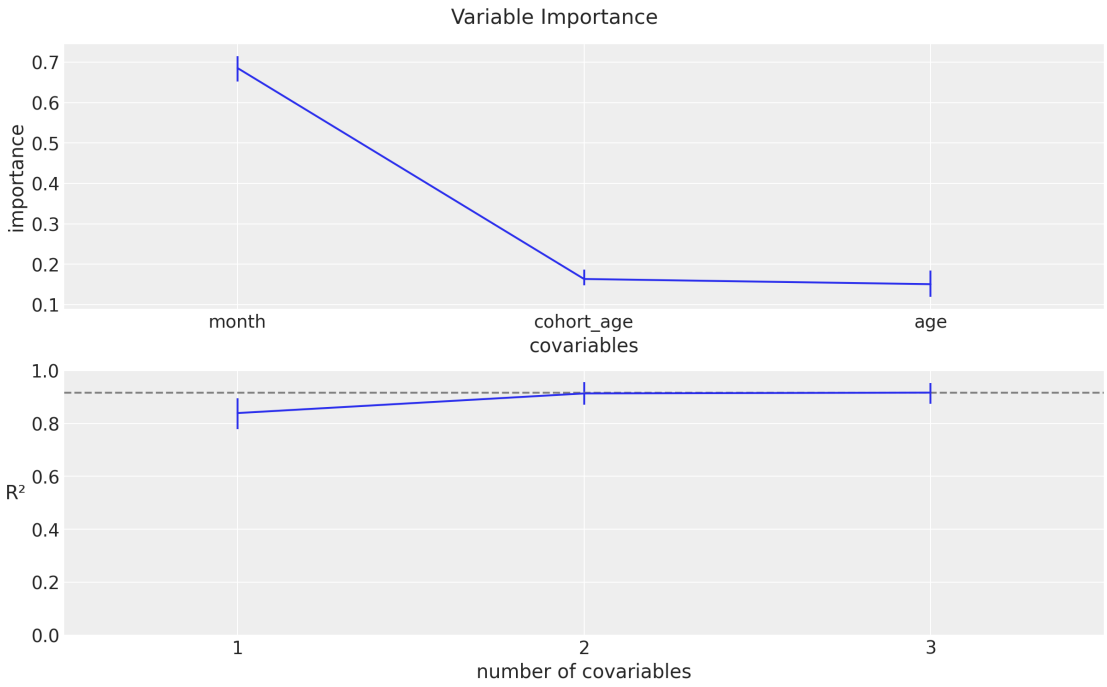

FIGURE 11. Variable importance for the retention component, quantified by each feature's contribution to in-sample R^2. This metric captures how much predictive power each feature adds to the model. Month (period) emerges as the most important feature, reflecting the strong seasonality observed in the exploratory analysis. Cohort age is also highly important, capturing the decay in retention as cohorts mature. The age (cohort identifier) has lower but non-negligible importance, suggesting some differences between cohorts beyond what is captured by cohort age and seasonality. These importance scores help prioritize data collection efforts and guide model simplification if needed, though all three features clearly contribute meaningfully to retention prediction.

The vertical dashed lines indicate the train/test split point. Several key observations emerge:

(1) The model successfully predicts both retention and revenue patterns for future periods, with most actual observations falling within the 94% HDI.
(2) The model effectively captures the seasonal patterns in retention, correctly predicting the expected peaks and troughs in future months based on historical patterns.
(3) For newer cohorts with limited training data (e.g., the *2022-07* cohort with only 4 data points in training), the model still produces reasonable predictions by leveraging information learned from older cohorts. This demonstrates effective transfer of knowledge across cohorts.

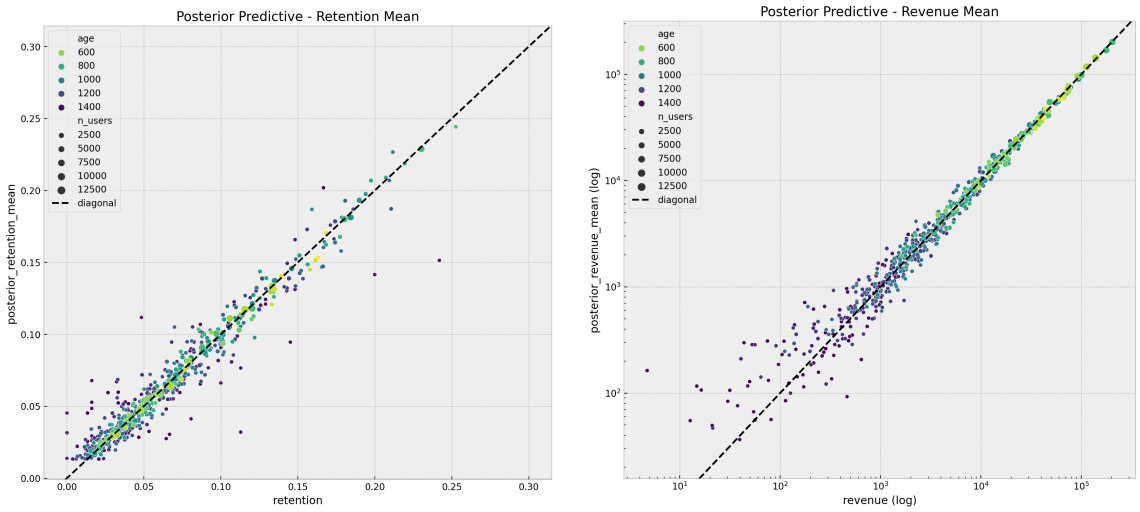

FIGURE 12. Retention (left) and revenue (right) in-sample posterior predictive mean values plotted against the actual observations. These scatter plots provide a quantitative assessment of model fit by comparing predicted versus observed values, with points closer to the diagonal line indicating better predictions.

(4) The 94% HDI appropriately widens for more distant future predictions, reflecting increasing uncertainty as we forecast further ahead.

These results highlight one of the key advantages of our Bayesian approach: the ability to make probabilistic forecasts with well-calibrated uncertainty using highest density intervals (HDI). The model provides not just point estimates but complete distributions, allowing businesses to understand the range of possible outcomes and make risk-aware decisions. The effective transfer of information across cohorts is particularly valuable for new cohorts where limited data is available.

4.3. **Business Value and Decision Impact.** Having demonstrated the model's predictive performance, we now discuss how these forecasts translate into business value and inform concrete decisions. The practical utility of any forecasting method ultimately depends on whether it enables better decision-making than available alternatives.
Probabilistic Forecasting for Risk Management. A key differentiator of our approach is the provision of complete posterior distributions rather than point estimates. Consider a marketing team planning acquisition budgets for the coming quarter. With point estimates alone, they might allocate resources based on expected customer lifetime value, but would lack information about downside risks. Our framework's 94% HDI provides explicit quantification of uncertainty: for instance, if a cohort's predicted 6-month revenue has a median of $50,000 but a 94% HDI of [$35,000, $68,000], decision-makers can assess whether the acquisition cost justifies this range of outcomes. Risk-averse organizations might base budgets on lower quantiles (e.g., 25th percentile) to ensure profitability

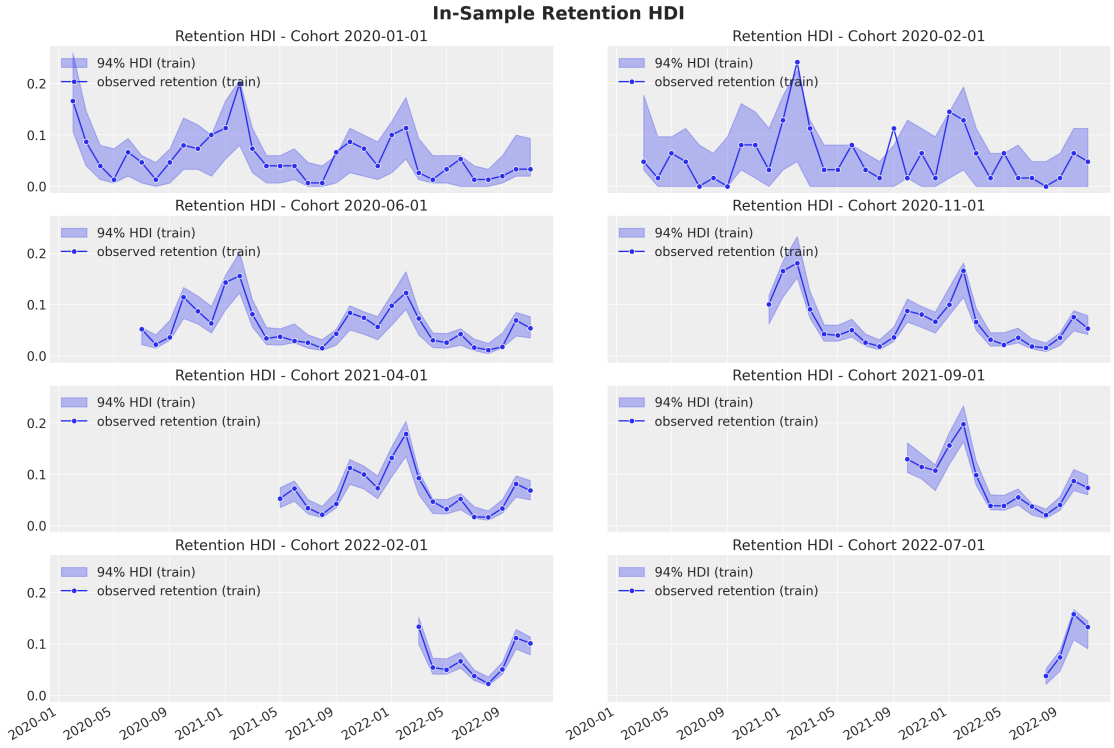

FIGURE 13. Retention in-sample posterior predictive distribution for selected cohorts, showing 94% HDI (blue shaded areas) and observed retention values (blue points). This visualization displays the model's predictive performance for retention across time for different cohorts, with uncertainty quantified through highest density intervals. The narrower intervals for more recent cohorts (bottom panels) reflect greater certainty due to more available data, while the consistent capture of observed values within the intervals indicates well-calibrated uncertainty estimates. The plots also reveal the model's ability to adapt to cohort-specific patterns and seasonal fluctuations, demonstrating its flexibility in capturing complex temporal dynamics.

even in pessimistic scenarios, while growth-focused companies might optimize for median expectations. Point estimates obscure this crucial risk dimension.

Seasonal Strategy Optimization. The model's explicit capture of seasonality enables month-specific strategic adjustments. The out-of-sample predictions reveal that retention peaks in November-December and dips mid-year—a pattern that persists across cohorts. Marketing teams can leverage this insight by timing customer acquisition campaigns to coincide with high-retention periods, maximizing long-term value per acquisition dollar. Conversely, they might implement targeted retention interventions during low-retention months to counteract natural attrition patterns. Financial planning can also benefit:

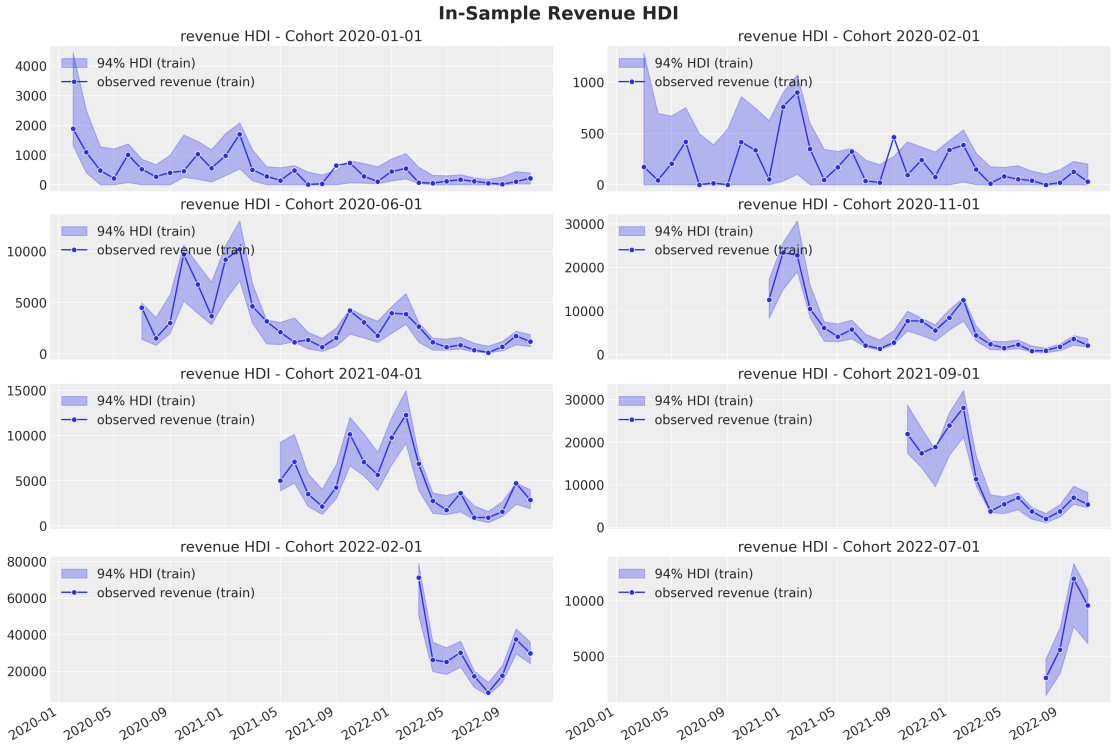

FIGURE 14. Revenue in-sample posterior predictive distribution for selected cohorts, showing 94% HDI (blue shaded areas) and observed revenue values (blue points). These plots illustrate the model's revenue predictions and associated uncertainty across time for different cohorts. The successful capture of observed values within the HDI bands demonstrates the model's ability to accurately represent not just central tendencies but also the inherent variability in revenue. The visualization highlights how our coupled modeling approach effectively propagates uncertainty from the retention component to revenue estimates, providing business stakeholders with realistic confidence intervals for financial planning and analysis.

quarterly revenue forecasts should account for seasonal fluctuations rather than assuming constant monthly patterns, avoiding systematic over- or under-estimation.

New Cohort Forecasting and Launch Planning. The information-sharing mechanism proves particularly valuable when evaluating new product launches or expansion into new markets. Consider a company launching a product variant or entering a new geographic region: the first cohort will have extremely limited data, but our model can generate reasonable forecasts by leveraging patterns from existing cohorts. While these predictions carry greater uncertainty (reflected in wider HDIs), they still provide actionable guidance for initial resource allocation. As data accumulates, forecasts automatically improve and

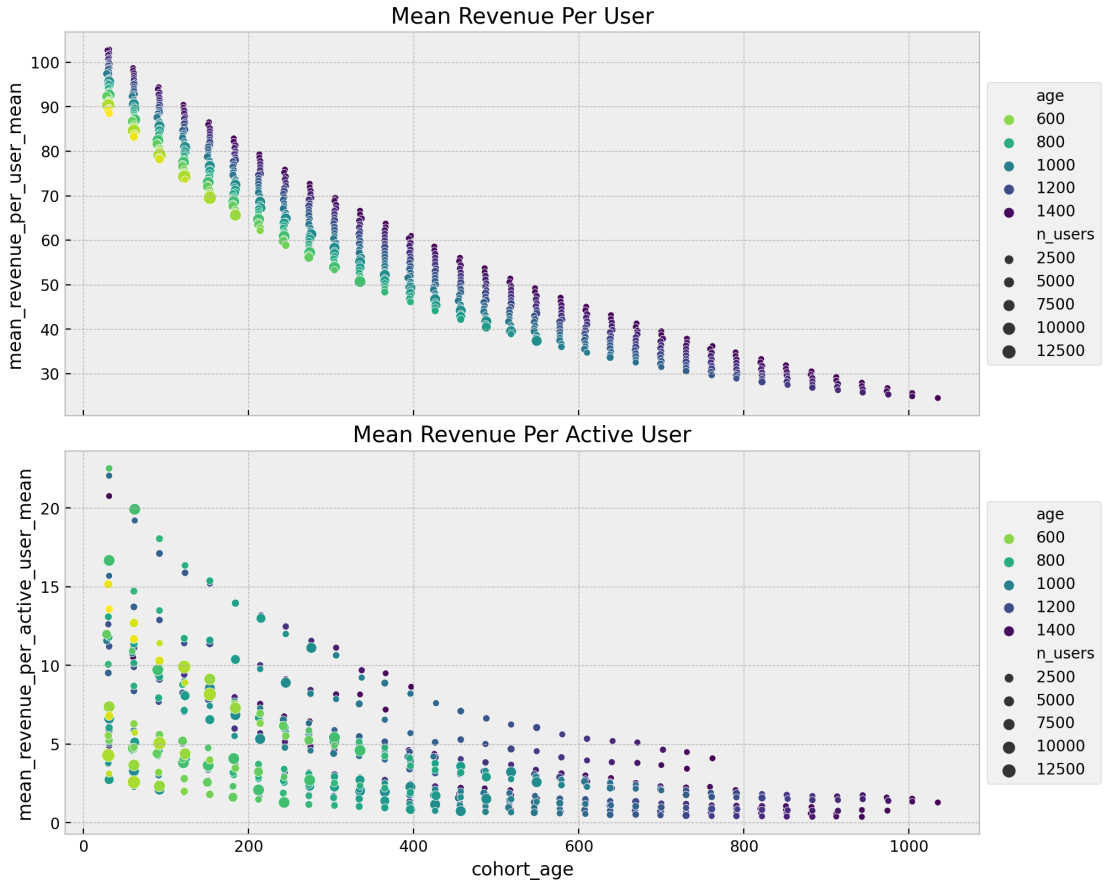

FIGURE 15. Additional view of posterior predictions across cohorts, illustrating the model's ability to capture cohort-specific patterns. This panel view organizes predictions by cohort (columns) and shows how the model adapts to the unique characteristics of each customer group.

uncertainty narrows—a dynamic that deterministic models cannot capture. This supports iterative decision-making: initial cautious investment based on high-uncertainty forecasts, followed by scaled investment as evidence accumulates.

Portfolio-Level Resource Allocation. Organizations managing multiple customer segments or product lines can use the framework to optimize resource allocation at the portfolio level. By generating forecasts for each cohort or segment, decision-makers can identify which groups merit increased retention investment versus which would benefit more from improved acquisition efficiency. For example, if older cohorts show declining revenue-per-active-user but stable retention, product enhancements targeting monetization might yield better returns than generic retention campaigns. Conversely, cohorts with strong monetization but declining retention would benefit from engagement initiatives. The model's coupled structure ensures that these strategic choices account for interdependencies between retention and revenue.

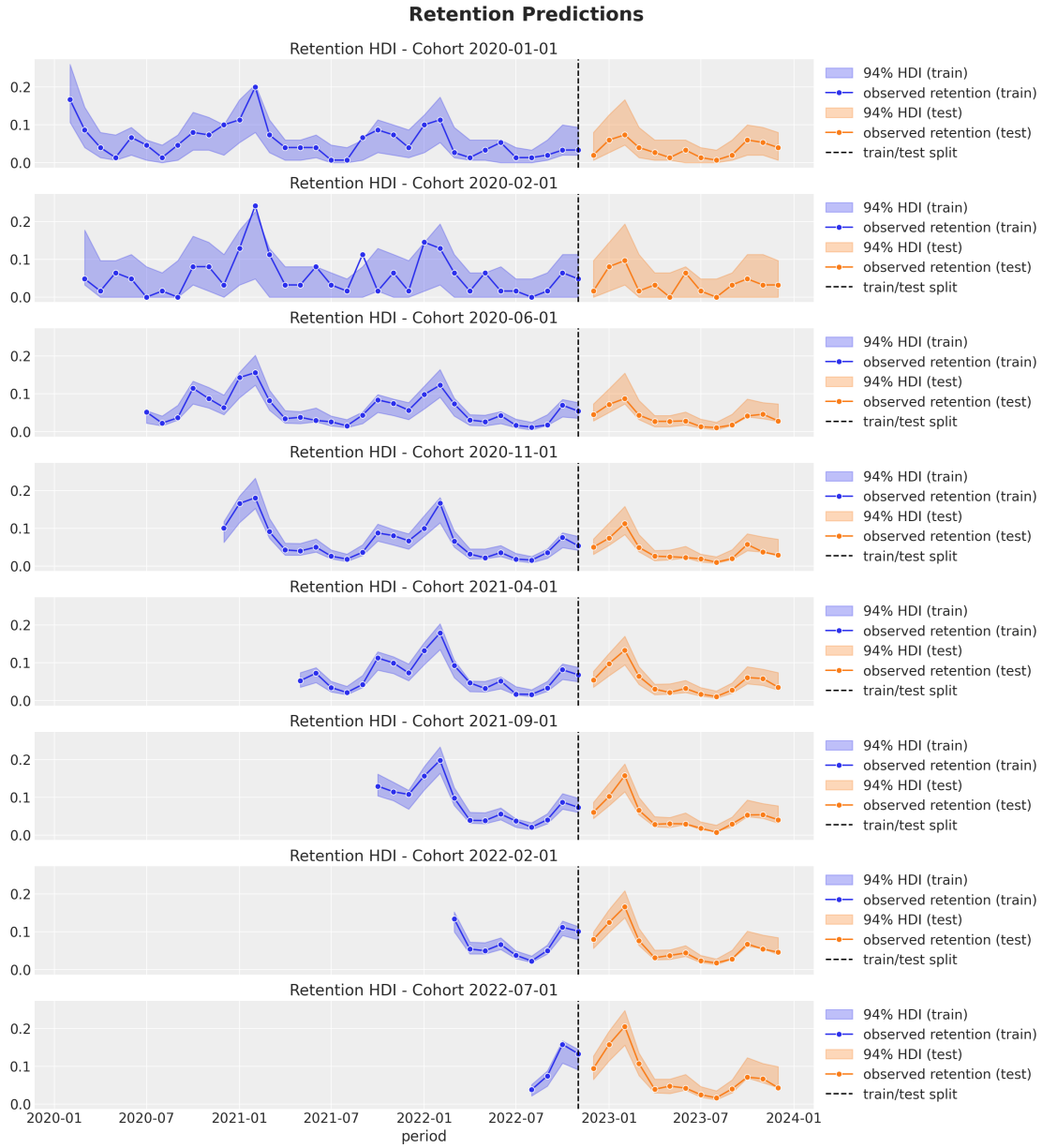

FIGURE 16. Retention out-of-sample posterior predictive distribution for (random) selected cohorts.

Threshold-Based Decision Rules. Many business decisions involve threshold rules: whether to continue a product line, expand to a region, or modify a pricing tier. Our probabilistic forecasts enable threshold-based analysis: what is the probability that a cohort's 12-month revenue exceeds the break-even point? For a new subscription tier, what is the probability retention exceeds the target of 60% at 6 months? These questions map

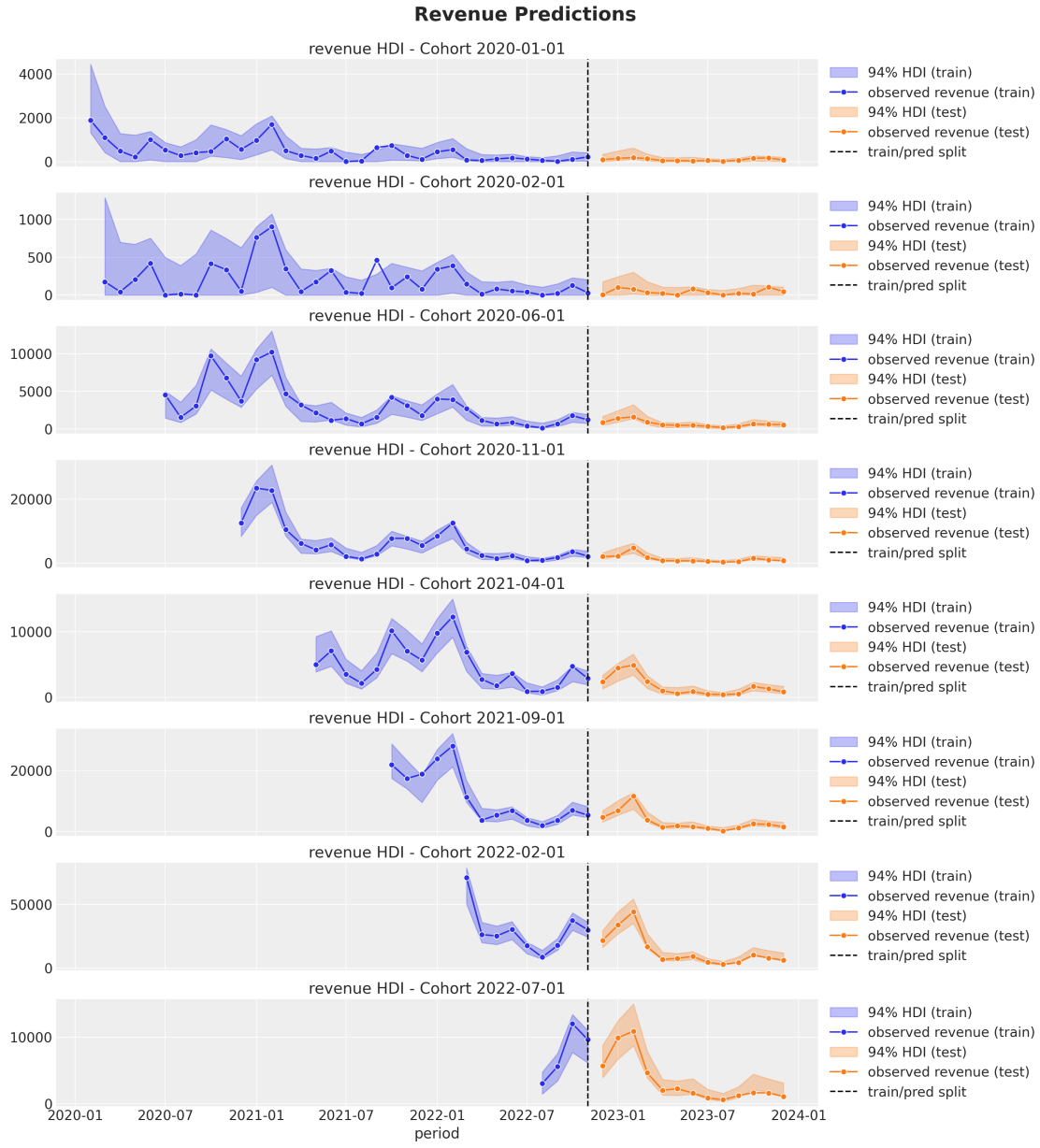

FIGURE 17. Revenue out-of-sample posterior predictive distribution for (random) selected cohorts.

naturally to posterior probabilities that our framework provides directly. Deterministic forecasts force organizations to treat inherently uncertain outcomes as binary decisions, potentially leading to premature termination of promising initiatives or continued investment in underperforming ones.

In summary, the business value of our approach extends beyond improved forecast accuracy (relative to simpler baselines, though we do not quantify this here). The combination of well-calibrated uncertainty, explicit seasonality, and cohort-level granularity enables a richer set of business decisions than point-estimate forecasting allows. Organizations can move from asking "what will happen?" to "what might happen, with what probabilities?"—a fundamental shift that supports more nuanced and risk-aware strategic planning.

5. Alternative Non-Parametric Approaches

The framework we have presented centers on BART for the retention component, a choice motivated by BART's flexibility, interpretability through PDP/ICE plots, and relatively straightforward hyperparameter tuning. However, the modular nature of our approach means the BART component can be replaced with other flexible function approximators. In this section, we briefly discuss one particularly promising alternative: neural networks with Bayesian inference. This discussion serves two purposes: first, it demonstrates the framework's flexibility and extensibility; second, it provides practitioners with guidance on when alternative implementations might be preferable.

While Bayesian Additive Regression Trees provide a powerful non-parametric approach for modeling the retention component, neural networks coupled with efficient Bayesian inference techniques offer an alternative that combines flexibility with computational efficiency, albeit with some tradeoffs in interpretability.

5.1. Neural Networks with NumPyro.

As demonstrated by [Orduz, 2024], the BART component in our model can be replaced with a neural network implemented using Flax, with inference performed using NumPyro [Phan et al., 2019]. The modified model structure becomes:

$$
\begin{aligned}
\text{Revenue} &\sim \text{Gamma}(N_{\text{active}}, \lambda) \\
\log(\lambda) &= (\text{intercept} \\
&\quad + \beta_{\text{cohort age}} \times \text{cohort age} \\
&\quad + \beta_{\text{age}} \times \text{age} \\
&\quad + \beta_{\text{cohort age}\times\text{age}} \times \text{cohort age} \times \text{age}) \\
N_{\text{active}} &\sim \text{Binomial}(N_{\text{total}}, p) \\
\text{logit}(p) &= \text{NN}(\text{cohort age}, \text{age}, \text{month})
\end{aligned}
$$

where NN represents a neural network. Even a simple architecture with one hidden layer containing just 4 units and sigmoid activation functions can capture the complex patterns in retention data effectively.

5.2. Advantages of the Neural Network Approach.

This neural network approach offers several advantages:

(1) **Computational efficiency**: Inference can be performed using stochastic variational inference (SVI), which is significantly faster than the MCMC sampling

required for BART models. This enables rapid model iteration and scaling to larger datasets.

(2) **Flexibility in inference methods**: Beyond SVI, the NumPyro framework allows for various sampling methods, including NUTS (No U-Turn Sampler) for full Bayesian inference when needed, as well as integration with other JAX-based probabilistic programming tools like BlackJax ([Cabezas et al., 2024]). To be fair, this can also be done with PyMC thanks to the PyTensor backend.

(3) **Comparable predictive performance**: Experiments on the same synthetic dataset show that the neural network approach produces similar retention and revenue predictions as the BART-based model, with well-calibrated 94% HDIs that appropriately capture uncertainty.

(4) **Development workflow**: The computational efficiency enables an iterative workflow where initial model development and testing can use fast SVI methods, with final inference performed using full MCMC sampling if desired.

5.3. **Limitations of Neural Networks Compared to BART.** Despite these advantages, the neural network approach does have some limitations when compared to BART:

(1) **Reduced interpretability**: Unlike BART, neural networks do not naturally provide partial dependence plots (PDP) or individual conditional expectation (ICE) plots. These visualizations, which help understand how individual predictors affect the target variable, require additional custom implementation with neural networks.

(2) **Architecture selection**: Neural networks require specification of the network architecture (number of layers, units per layer, activation functions), which introduces additional hyperparameters that must be selected, whereas BART requires fewer tuning decisions.

5.4. **Practical Considerations.** The choice between BART and neural network approaches depends on the specific needs of the application:

- For applications where interpretability is paramount and computational efficiency is less critical, BART may be preferred.
- For large-scale applications where inference speed is essential or when rapid model iteration is needed, the neural network approach with SVI offers significant advantages.
- In some cases, a hybrid approach might be valuable—using the faster neural network model for initial exploration and prototyping, then moving to BART for final analysis when interpretability is needed.

The implementation details and complete code examples for the neural network approach can be found in [Orduz, 2024].

6. Extension to Hierarchical Multi-Market Modeling

6.1. **Motivation and Approach.** Organizations operating across multiple markets or customer segments frequently encounter an asymmetry in data availability: some markets are mature with extensive cohort histories, while others are nascent with only a few

observed cohorts. Modeling each market independently wastes valuable information that could be shared across markets, while complete pooling ignores market-specific dynamics. A hierarchical structure provides an elegant solution to this challenge by enabling information pooling while preserving market-specific patterns.

The business motivation for hierarchical modeling is compelling. Consider a company expanding into new geographic regions or launching products in new market segments. Early-stage markets lack the data needed for reliable independent forecasts, yet business decisions—resource allocation, growth projections, strategic planning—cannot wait years for sufficient data accumulation. By borrowing information from more established markets through hierarchical priors, we can generate credible forecasts for young markets that would otherwise be impossible to model reliably.

We extend the neural network approach from Section 5 to accommodate multiple markets through two key modifications:

- **Retention component**: We incorporate market identity as an additional feature in the neural network. The network learns market-specific retention patterns while sharing information about temporal dynamics (cohort age, calendar effects, seasonality) across markets.
- **Revenue component**: We implement a hierarchical linear model where market-specific regression coefficients are drawn from common hierarchical priors. This allows each market to have its own revenue dynamics while constraining these parameters to be similar across markets through the prior distribution.

The hierarchical structure naturally addresses the data asymmetry problem: markets with abundant data inform the hierarchical priors, which in turn regularize predictions for data-sparse markets toward sensible values. Crucially, the coupling mechanism between retention and revenue components remains intact in this hierarchical setting—the number of active users predicted by the retention model still informs the revenue model's shape parameter.

6.2. **Synthetic Multi-Market Data.** To demonstrate the hierarchical extension, we extend our synthetic data generation process to create cohort-level observations across four markets with varying maturity levels (recall we train until November 2022 in the synthetic data generation process):

- **Market A**: Starting from January 2020 (mature, most data)
- **Market B**: Starting from February 2021 (moderately mature)
- **Market C**: Starting from January 2022 (developing)
- **Market D**: Starting from July 2022 (youngest)

The data generation process maintains the same retention and revenue dynamics described in Section 2, but now each market has its own data realization. We apply the same train/test split strategy, holding out the most recent periods for validation. Figure 18 visualizes the data structure, showing cohort availability and revenue patterns across the four markets.

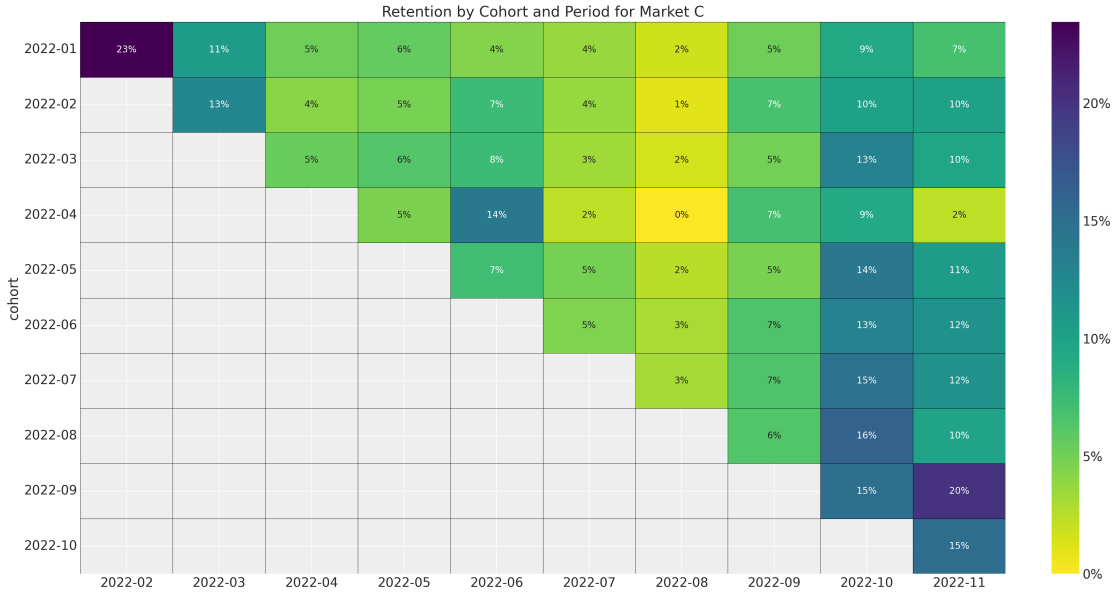

FIGURE 18. Retention matrix for Market C. We just have 10 cohorts of data available for this market.

This multi-market setup creates a realistic challenge: can we forecast revenue for Market C, which has limited data, by leveraging patterns learned from Markets A, B, and D?

6.3. **Results and Information Pooling.** The hierarchical model achieves strong predictive performance across all markets. Figure 19 shows revenue predictions for Market C by leveraging patterns learned from Markets A, B, and D, demonstrating that the model successfully borrows strength from more mature markets to generate accurate forecasts despite having only 10 cohorts.

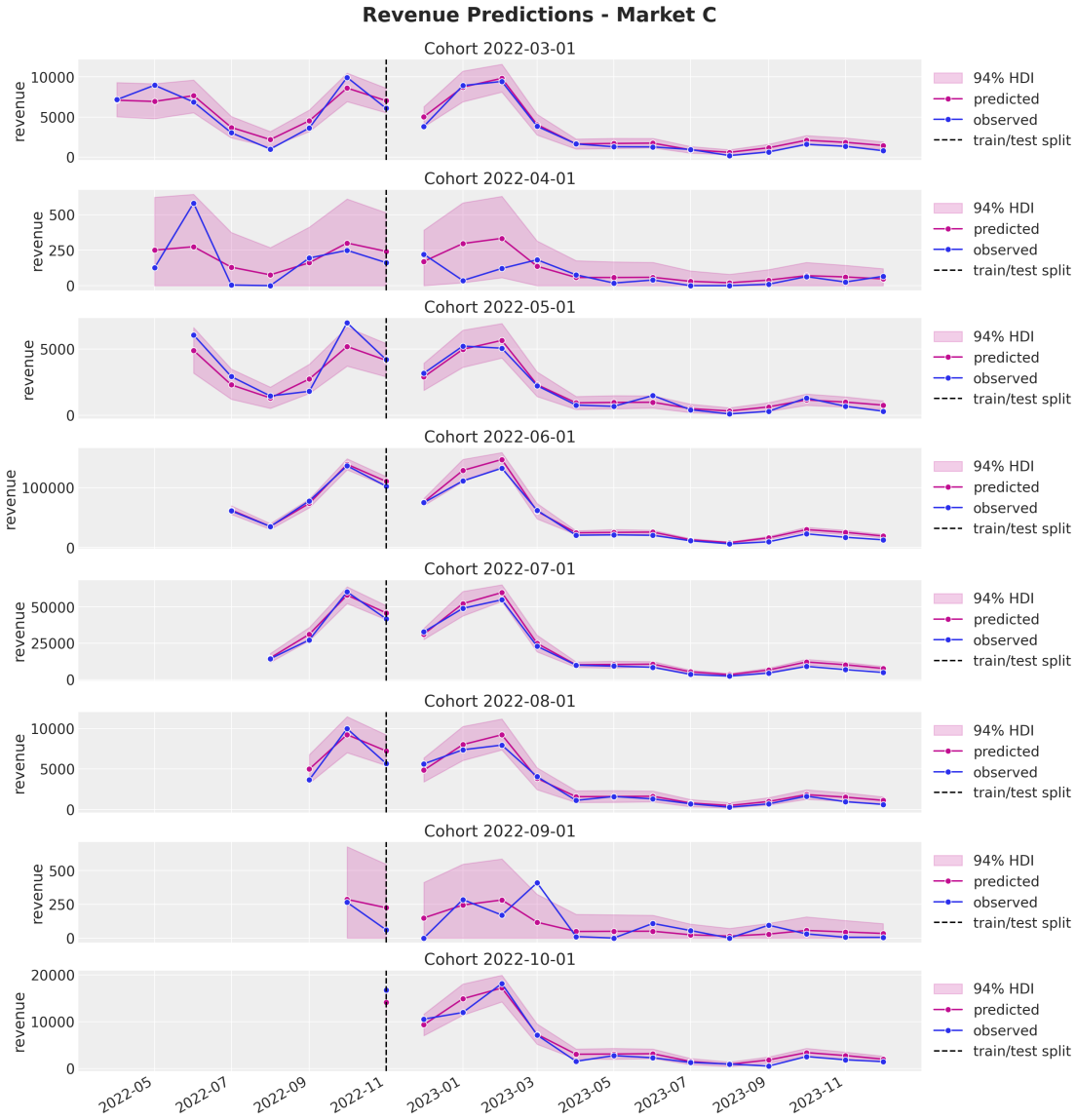

FIGURE 19. Revenue predictions for four selected cohorts in Market C.

Several key findings emerge from the hierarchical implementation:

- **Successful information pooling**: Market D achieves prediction accuracy that would be impossible with independent modeling. The hierarchical priors effectively transfer knowledge about retention decay patterns and revenue dynamics from mature markets.
- **Market-specific adaptation**: While borrowing information, the model still captures market-specific patterns. Each market's predictions reflect its own data when available, rather than being dominated by the larger markets.

- **Appropriate uncertainty quantification**: The highest density intervals are wider for Market D than for more established markets, correctly reflecting the greater uncertainty due to limited data. This honest uncertainty quantification is crucial for business decision-making.
- **Coupling mechanism preserved**: The connection between retention and revenue components functions effectively in the hierarchical setting. Active user predictions from the retention model inform revenue forecasts across all markets.

The hierarchical extension demonstrates that the framework's core architecture—the coupling between retention and revenue— naturally extends to more complex settings. This extensibility is not accidental but rather a consequence of the modular design: the coupling mechanism operates at the cohort-observation level and is agnostic to whether those observations come from a single market or multiple markets with hierarchical structure.

6.4. **Implementation Notes.** For computational efficiency with multiple markets, we implement the hierarchical model using Stochastic Variational Inference (SVI) in NumPyro rather than MCMC sampling. SVI provides approximate posterior distributions through optimization, enabling the model to scale to tens or hundreds of markets where full MCMC would be computationally prohibitive. For applications with fewer markets (say, 5-10), MCMC remains a viable alternative that may provide more accurate uncertainty quantification.

The hierarchical priors on revenue model parameters follow standard conventions: market-specific coefficients are drawn from normal distributions whose means and variances are themselves parameters with hyperpriors. The neural network component for retention incorporates market identity through one-hot encoding, allowing the network to learn market-specific functions while sharing the learned representations across markets.

The complete implementation, including data generation, model specification, and visualization code, is available at [Orduz, 2025]. The implementation demonstrates that extending the base framework to hierarchical structures requires modest additional complexity—primarily the specification of hierarchical priors and the inclusion of market identifiers in the feature set.

7. Conclusion and Future Directions

7.1. **Broader Implications and Future Directions.** The ability to accurately forecast retention and revenue metrics represents a significant competitive advantage in today's business environment. Having addressed practical considerations, we now turn to broader implications of this work. In this paper, we have presented a Bayesian framework for jointly modeling cohort-level retention and revenue that addresses critical gaps in the existing literature on customer lifetime value and cohort analysis. As discussed in our review of related work, traditional approaches—whether individual-level BTYD models, linear age-period-cohort frameworks, or single-outcome survival models—face limitations in flexibility, computational scalability, joint modeling of related outcomes, and principled uncertainty quantification. Our contribution directly addresses these gaps through a coupled retention-revenue architecture that balances sophisticated non-parametric modeling with practical interpretability.

By combining the flexibility of Bayesian additive regression trees with the interpretability of linear models, our approach offers both analytical power and practical utility. The choice to operate at the cohort level is not a limitation but rather a deliberate strategic decision that aligns with how many business decisions are made, reduces noise through aggregation, and provides computational efficiency—while still allowing for the incorporation of rich covariate information when needed. Our framework provides several distinctive advantages that merit highlighting:

(1) **Adaptive complexity**: The BART component automatically adjusts its complexity to match the underlying patterns in the retention data, capturing nonlinear relationships and interactions that would be difficult to specify manually. Meanwhile, the linear component for revenue provides clear interpretability of key drivers, offering the best of both worlds—sophisticated modeling where needed and transparency where possible.
(2) **Principled uncertainty quantification**: Unlike deterministic approaches that provide only point estimates, our Bayesian framework generates complete posterior distributions for all quantities of interest. This allows decision-makers to understand the full range of potential outcomes through 94% highest density intervals (HDI) and tailor their strategies to their risk preferences. For instance, a risk-averse business might base resource allocation decisions on lower quantiles of the revenue prediction distribution rather than mean estimates.
(3) **Knowledge transfer across cohorts**: The model's structure enables effective information sharing between cohorts, leveraging patterns from data-rich older cohorts to improve predictions for newer cohorts with limited history. This is particularly valuable in fast-growing businesses where the latest cohorts often represent significant portions of the customer base yet have the least historical data.
(4) **Customizable architecture**: The modular design allows for straightforward extensions to incorporate business-specific factors and external variables. Whether integrating marketing channel information, product usage metrics, or macroeconomic indicators, the model can adapt to diverse business contexts without fundamental redesign. As demonstrated in Section 6, the framework naturally extends to hierarchical structures for multi-market modeling, where information pooling across markets enables reliable forecasts even for data-sparse segments—a critical capability for growing businesses operating across multiple geographies or customer segments.

Our experiments with synthetic data demonstrate the model's effectiveness, but the real value of this approach emerges in practical business applications. By providing both accurate forecasts and well-calibrated uncertainty estimates through highest density intervals (HDI), this methodology enables more informed decision-making across multiple business functions:

- **Marketing teams** can optimize acquisition spending based on expected customer lifetime value, potentially varying their strategies seasonally based on predicted retention patterns.

- **Product teams** can prioritize features that target high-value cohorts or address specific drop-off points in the customer lifecycle.
- **Financial planning** becomes more robust with probabilistic forecasts that account for the inherent uncertainty in future customer behavior.
- **Customer success initiatives** can be tailored to specific cohorts based on their predicted retention trajectories, potentially intervening at critical points to improve outcomes.

Despite these advantages, we acknowledge some limitations and boundary conditions that present opportunities for future research. First, while our cohort-level approach offers strategic value and practical advantages as discussed earlier, it naturally does not provide individual-level predictions. Organizations requiring customer-specific forecasts for personalization or targeting would need to complement this approach with individual-level models—though as we have argued, many strategic business decisions operate precisely at the cohort level where our framework excels. Second, the current framework assumes that cohort behavior patterns remain relatively stable over time, with seasonal variations occurring around consistent trends. In rapidly evolving markets or during significant disruptions (e.g., major product changes, economic shocks), this assumption may not hold. Future work could explore regime-switching models or online learning approaches that adapt more quickly to fundamental shifts in customer behavior. Third, while we have demonstrated that individual-level covariates can be incorporated into the BART component, our synthetic data examples use only temporal features. Empirical validation with real datasets incorporating rich customer characteristics—acquisition channel, geographic location, demographic attributes—would provide additional evidence of the framework's flexibility and the scalability advantages of BART with many features.

Looking ahead, several promising research directions emerge:

(1) **Causal modeling extensions**: Incorporating causal inference techniques to estimate the impact of interventions on retention and revenue would enhance the model's utility for decision support.
(2) **Multi-product ecosystems**: Extending the framework to handle customers who engage with multiple products or services, capturing cross-product effects on retention and spending.
(3) **Advanced hierarchical structures**: While Section 6 demonstrates hierarchical modeling across markets, further extensions could address more complex nested structures (e.g., markets within regions, products within categories) or time-varying hierarchical parameters that adapt to evolving market dynamics.

The methodology presented in this paper represents a significant step forward in cohort-based retention and revenue modeling, addressing specific gaps in the age-period-cohort, survival analysis, and customer lifetime value literatures through principled Bayesian coupling of retention and revenue outcomes. By embracing the complexity inherent in customer behavior while maintaining analytical tractability and interpretability, our approach bridges the gap between sophisticated statistical techniques and practical business applications. The framework's flexibility—demonstrated through both BART and neural network implementations—combined with its focus on cohort-level strategic

insights positions it as a valuable complement to existing individual-level CLV models. As companies continue to recognize the strategic importance of customer retention and lifetime value, flexible and accessible modeling approaches like the one presented here will become increasingly essential tools in the modern business analytics toolkit, particularly for strategic decision-making contexts where aggregate patterns and well-calibrated uncertainty estimates drive resource allocation and planning.

References

[Abril-Pla et al., 2023] Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T., and Zinkov, R. (2023). Pymc: A modern and comprehensive probabilistic programming framework in python. *PeerJ Computer Science*, 9:e1516.

[Cabezas et al., 2024] Cabezas, A., Corenflos, A., Lao, J., and Louf, R. (2024). BlackJAX: Composable Bayesian inference in JAX.

[Chipman et al., 2010] Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

[Fader et al., 2005] Fader, P., Hardie, B., and Lee, K. (2005). "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science*, 24:275–284.

[Fader and Hardie, 2007a] Fader, P. S. and Hardie, B. G. (2007a). How to project customer retention. *Journal of Interactive Marketing*, 21(1):76–90.

[Fader and Hardie, 2007b] Fader, P. S. and Hardie, B. G. (2007b). Incorporating Time-Invariant Covariates into the Pareto/NBD and BG/NBD Models. `http://brucehardie.com/notes/019/`.

[Fader and Hardie, 2017] Fader, P. S. and Hardie, B. G. (2017). Fitting the sBG Model to Multi-Cohort Data. `http://brucehardie.com/notes/017/`.

[Fannon and Nielsen, 2018] Fannon, Z. and Nielsen, B. (2018). Age-Period-Cohort Models. Technical Report 2018-W04, Nuffield College, University of Oxford.

[Hubbard et al., 2021] Hubbard, D., Rostykus, B., Raimond, Y., and Jebara, T. (2021). Beta Survival Models. In *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*, volume 146 of *Proceedings of Machine Learning Research*, pages 22–39. PMLR.

[Orduz, 2022] Orduz, J. (2022). A Simple Cohort Retention Analysis in PyMC. `https://juanitorduz.github.io/retention/`.

[Orduz, 2023a] Orduz, J. (2023a). Cohort Retention Analysis with BART. `https://juanitorduz.github.io/retention_bart/`.

[Orduz, 2023b] Orduz, J. (2023b). Cohort Revenue & Retention Analysis: A Bayesian Approach. `https://juanitorduz.github.io/revenue_retention/`.

[Orduz, 2023c] Orduz, J. (2023c). Cohort Revenue & Retention Analysis: A Bayesian Approach - Code to generate data. `https://github.com/juanitorduz/website_projects/blob/master/Python/retention_data.py`.

[Orduz, 2024] Orduz, J. (2024). Cohort Revenue Retention Analysis with Flax and NumPyro. `https://juanitorduz.github.io/revenue_retention_numpyro/`.

[Orduz, 2025] Orduz, J. (2025). Hierarchical Revenue & Retention Modeling. `https://juanitorduz.github.io/hierarchical_revenue_retention/`.

[Phan et al., 2019] Phan, D., Pradhan, N., and Jankowiak, M. (2019). Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv preprint arXiv:1912.11554*.

[PyMC-Labs, 2023] PyMC-Labs (2023). PyMC-Marketing: Bayesian marketing toolbox in PyMC. `https://github.com/pymc-labs/pymc-marketing`. Media Mix (MMM), customer lifetime value (CLV), buy-till-you-die (BTYD) models and more.

[Quiroga et al., 2022] Quiroga, M., Garay, P. G., Alonso, J. M., Loyola, J. M., and Martin, O. A. (2022). Bayesian additive regression trees for probabilistic programming.

[Stucchio, 2015] Stucchio, C. (2015). Bayesian a/b testing at vwo. `https://vwo.com/downloads/VWO_SmartStats_technical_whitepaper.pdf`.

Email address: juanitorduz@gmail.com
URL: https://juanitorduz.github.io/

Berlin, Germany