COHORT REVENUE & RETENTION ANALYSIS: A BAYESIAN APPROACH

JUAN CAMILO ORDUZ AND DANIEL GUHL

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.

Contents

1. Introduction

Understanding and predicting customer behavior directly impacts business profitability through improved retention strategies and resource allocation. Among the metrics

Date: February 3, 2026.

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., 2005a], 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 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

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

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

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 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 a significant amount of work to implement in a way that scales).

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).

Visualizing Cohort Data. To motivate the modeling approach, we first illustrate the type of data structures our framework is designed to handle. Figure 1 shows a typical cohort-level retention matrix. These cohorts are typically defined by business rules, e.g., customers who registered, downloaded the app, or made their first purchase during the same month. Older cohorts have more historical data, and we would like to leverage this information to improve estimation for younger cohorts. That is, we do not want to model each cohort independently but rather the *whole retention matrix*.

Similarly, we want to model the corresponding revenue matrix (Figure 2), ensuring we use all available information to improve revenue estimation for younger cohorts. A key observation is the strong correlation between revenue and the number of active users (Figure 3), which motivates our coupling mechanism.

The rest of this paper is organized as follows. Section 2 reviews related work in age-period-cohort modeling, survival analysis, and customer lifetime value. Section 3 presents our core theoretical contribution: a coupled retention-revenue framework where the number of active users links both components. Section 4 instantiates the framework using synthetic data, including model diagnostics and in-sample/out-of-sample predictive performance evaluation. Section 5 discusses alternative non-parametric approaches using neural networks. Section 6 extends the framework to hierarchical multi-market settings. Section 7 validates the approach on real-world H&M transaction data, followed by a discussion of model limitations and trade-offs in Subsection 7.2. Finally, Section 8 concludes.

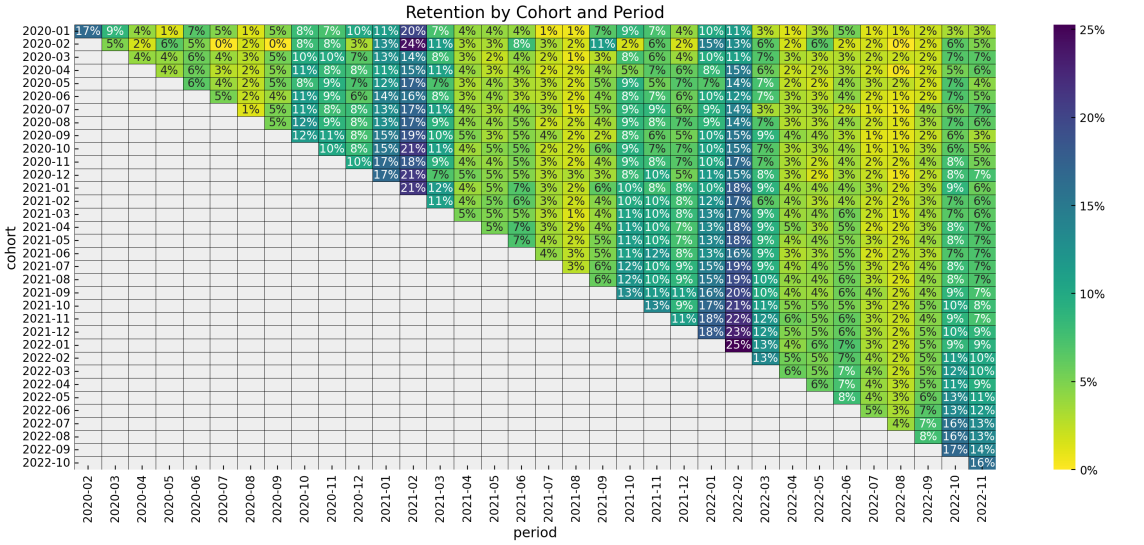

FIGURE 1. Retention matrix example. The matrix visualizes customer retention rates across different cohorts (rows) and observation periods (columns). Each cell represents the proportion of customers from a specific acquisition cohort that remained active in a subsequent period. Colors indicate retention 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.

2. Related Work

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.

2.1. **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

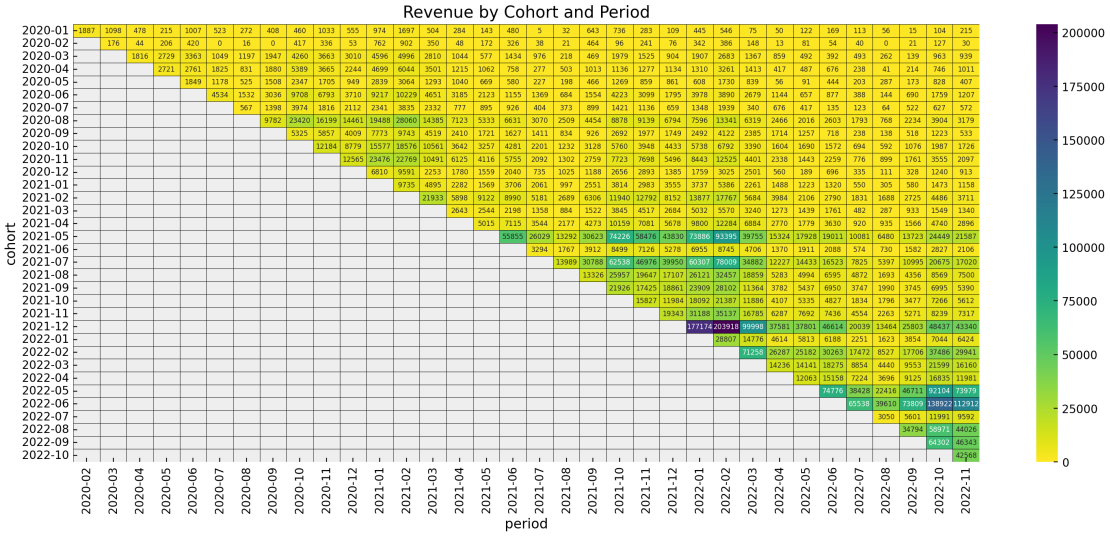

FIGURE 2. Revenue per cohort. This heatmap visualizes the total revenue generated by each cohort (rows) across different time periods (columns). The color intensity corresponds to revenue magnitude, revealing a strong correlation with the number of active users (Figure 3).

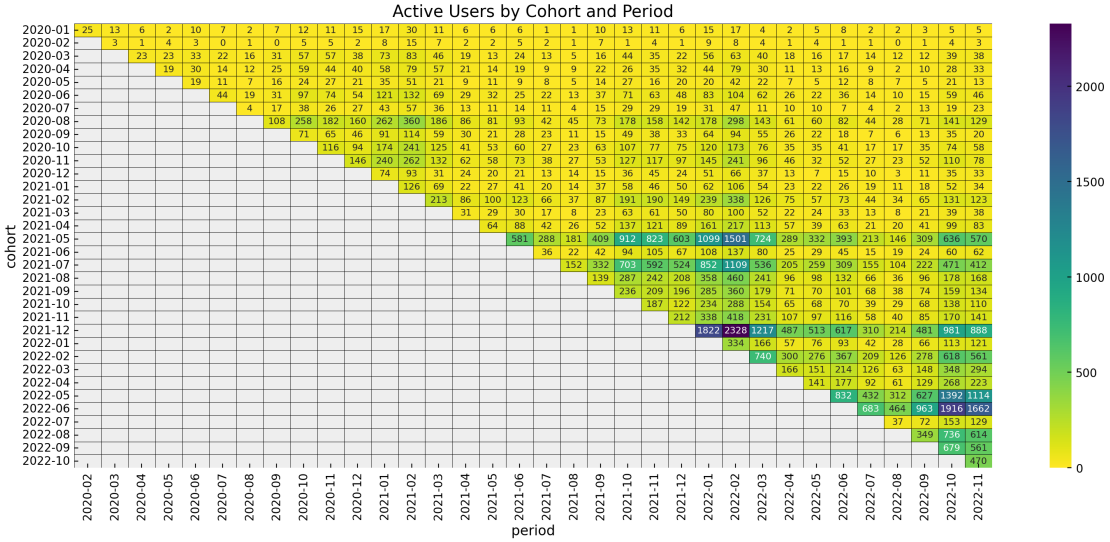

FIGURE 3. Number of active users across cohorts. This heatmap displays the absolute count of active users for each cohort (rows) across observation periods (columns).

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 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.

2.2. **Survival Analysis and Retention Modeling.** Recently, [Hubbard et al., 2021] introduced Beta Survival Models that use non-parametric methods 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 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.

2.3. **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., 2005a, 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 (Shifted Beta Geometric model), these extensions still work within the constraints of the original parametric model structures. We want to emphasize that we are not claiming that our approach is better than the BTYD models, but rather that in many applications where the seasonal components and additional covariates are important, for the mid-term decision making, our approach tends to be more accurate and flexible.

3. The Model Framework

This section presents the core theoretical contribution of this work: a Bayesian framework that couples retention and revenue through the number of active users. We first introduce the coupling mechanism, then describe the features used in the model, and finally discuss the flexibility in modeling the latent variables.

3.1. **The Core Coupling Mechanism.** The central insight of our approach is a principled coupling between retention and revenue through the number of active users. Let $i \in \{1, \ldots, I\}$ denote the acquisition cohort and $j \in \{0, \ldots, J_i\}$ denote the cohort age (time since acquisition). We model:

$$\boxed{\begin{aligned} R_{i,j} &\sim \text{Gamma}(N_{\text{active},i,j}, \lambda_{i,j}) \\ N_{\text{active},i,j} &\sim \text{Binomial}(N_{\text{total},i}, p_{i,j}) \end{aligned}}$$

where $p_{i,j}$ represents the retention probability and $\lambda_{i,j}$ is the rate parameter for revenue. This coupling extends decomposition ideas similar to those in the Gamma-Gamma model [Fader et al., 2005b], which separates transaction frequency from monetary value. The key insight is that the number of active users $N_{\text{active},i,j}$ a latent quantity we estimate through the retention model directly informs the revenue model as its shape parameter.

This formulation offers several important properties:

- **Consistent uncertainty propagation**: Uncertainty in retention estimates automatically propagates to revenue predictions through N_{active}.
- **Natural variance scaling**: The gamma parametrization ensures that cohorts with more active users have lower relative variance (coefficient of variation $1/\sqrt{N_{\text{active},i,j}}$), matching the intuition that aggregated revenue from larger cohorts should be more stable.
- **Interpretable mean structure**: The mean of the gamma distribution is $N_{\text{active},i,j}/\lambda_{i,j}$, allowing us to interpret $1/\lambda_{i,j}$ as the average revenue per active user for cohort i at age j.

Figure 4 shows the complete model structure in plate notation (common in Bayesian graphical models).

3.1.1. *Numerical Evidence: Uncertainty Propagation.* To demonstrate the importance of the coupling mechanism for uncertainty quantification, we compare two model variants: a *non-coupled* model that uses observed active user counts directly, and the *coupled* model that propagates uncertainty through the estimated N_{active} from the retention component.

Figures 5 and 6 reveal a critical difference in uncertainty quantification. The non-coupled model (Figure 5) produces HDIs that are systematically too narrow, particularly for younger cohorts where sample sizes are smaller. When revenue values approach zero, the non-coupled model's intervals frequently fail to contain the observed values, indicating poor calibration. In contrast, the coupled model (Figure 6) produces wider, more realistic HDIs that maintain proper coverage across all cohorts. This demonstrates that the coupling mechanism not only provides a principled connection between retention and revenue but also ensures that uncertainty in retention estimates appropriately propagates to revenue predictions.

3.2. **Features: Cohort Age and Period.** Before going into the details of modeling the latent variables p and λ, we define the core features used in the model. 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. It is natural to consider the following features to model the retention and revenue matrices:

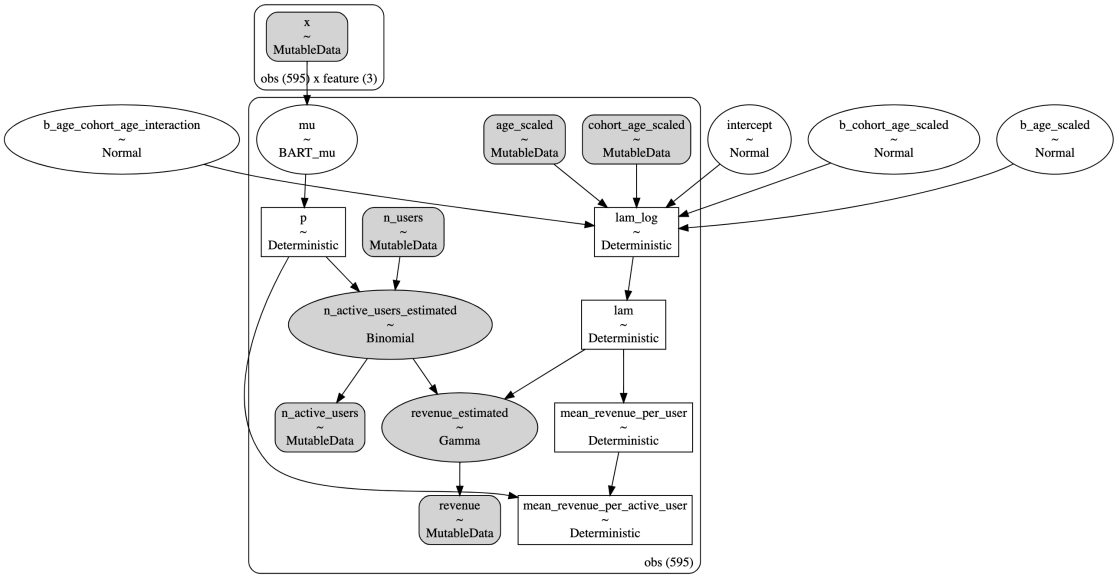

FIGURE 4. Cohort-revenue-retention model structure. This diagram illustrates the coupling mechanism that connects the two components of our framework. On the left, the retention component models the number of active users N_{active} as a binomial random variable, where the retention probability p is modeled as a function of 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.

- **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, additional covariates can be added to the model. The only requirement for out-of-sample predictions is that these covariates must be available for future observation periods.

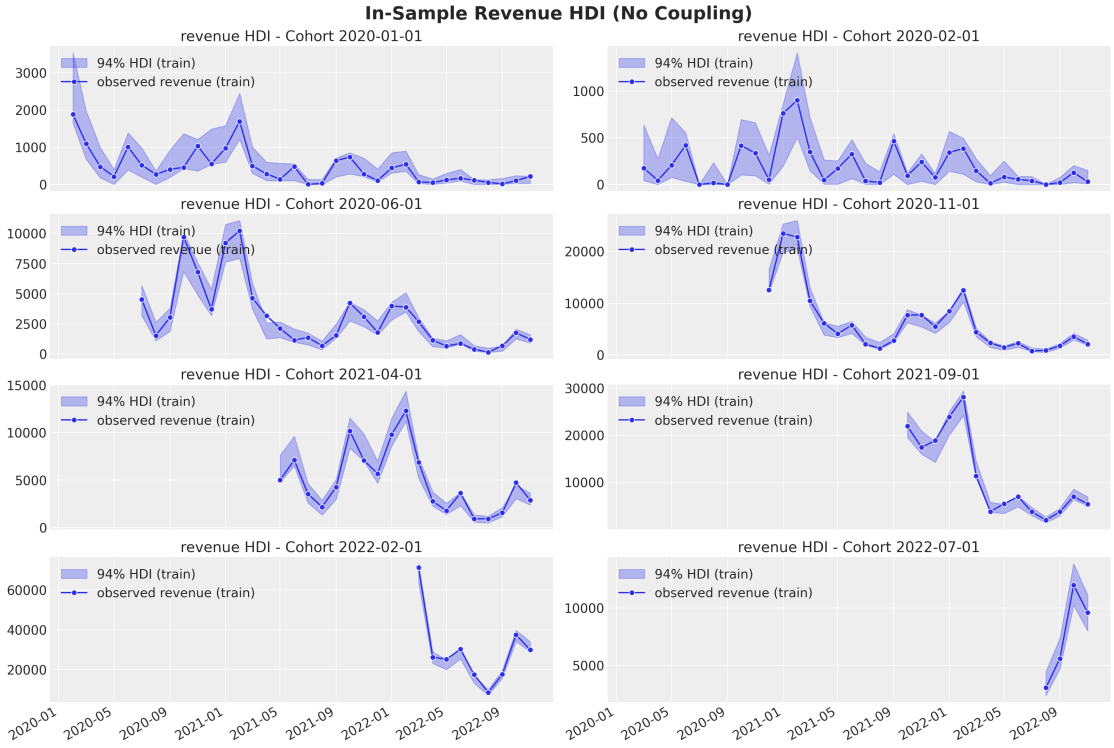

FIGURE 5. Revenue 94% HDI for the non-coupled model using observed active user counts. Younger cohorts (bottom panels) with smaller sample sizes show HDIs that are too narrow and fail to capture observed values near zero.

3.3. **Flexible Modeling of Latent Variables.** A key strength of our framework is the flexibility in how we model the latent variables p (retention probability) and λ (revenue rate). The coupling mechanism remains the same regardless of the specific functional forms chosen.

3.3.1. *Retention Probability p.* We recommend modeling the retention probability using flexible non-parametric methods that can capture complex relationships without requiring explicit specification of functional forms:

$$\text{logit}(p) = f(\text{cohort age}, \text{age}, \text{month})$$

where f can be:

- **BART (Bayesian Additive Regression Trees)**: Our recommended baseline. BART [Chipman et al., 2010] represents the unknown function as a sum of regression trees, automatically learning non-linear relationships and interactions. The PyMC implementation [Quiroga et al., 2022] provides interpretability tools such as partial dependence plots (PDP) and individual conditional expectation (ICE) plots.

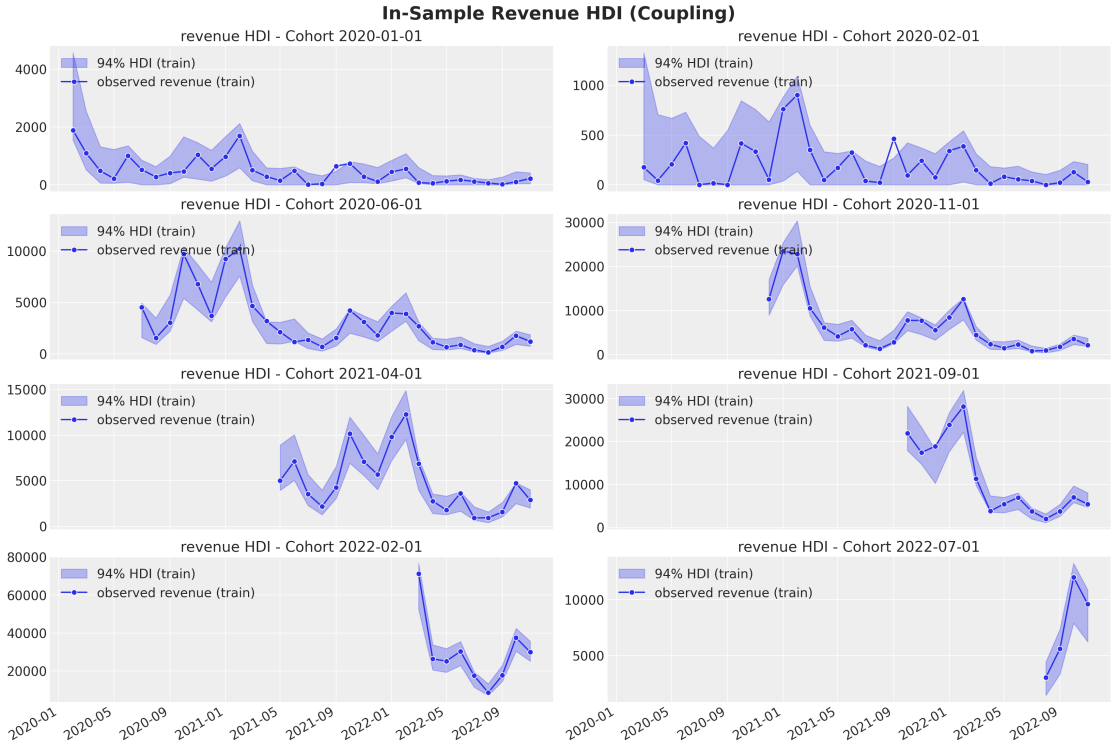

FIGURE 6. Revenue 94% HDI for the coupled model with uncertainty propagation through N_{active}. The wider HDIs provide better coverage across all cohorts, demonstrating proper uncertainty propagation.

- **Neural Networks**: For applications requiring fast inference or scaling to very large datasets, neural networks with stochastic variational inference offer an alternative (see Section 5).
- **Linear Models**: For simpler applications or as a baseline comparison, a linear model with interactions can be used.

3.3.2. *Revenue Rate λ.* For the revenue component, we typically find that simpler parametric models suffice:

$$\begin{aligned}\log(\lambda) = (&\text{intercept} \\ &+ \beta_{\text{cohort age}} \times \text{cohort age} \\ &+ \beta_{\text{age}} \times \text{age} \\ &+ \beta_{\text{cohort age}\times\text{age}} \times \text{cohort age} \times \text{age})\end{aligned}$$

A key insight from both synthetic and 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 through the coupling mechanism. However, if the business model has a strong seasonal component in average basket sizes, an additional seasonal term can be added.

3.4. **Advantages of the Framework.** This modeling strategy offers several distinct advantages:

- **Flexibility in relationship modeling**: The non-parametric 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 (e.g., macroeconomic indicators, marketing campaign intensities, etc.) can be seamlessly integrated into the model. The framework also extends naturally to hierarchical multi-market settings (Section 6).
- **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.

Remark 1 (Conditioning on Active Users)**.** When running inference (MCMC or SVI), we condition on the number of active users N_{active} rather than directly on the observed retention rates. The reason to treat retention as a latent variable is to obtain better uncertainty estimates depending on the cohort size (N_{total}) and the number of active users (N_{active}). Recall that the standard error of a proportion is $\sqrt{p(1-p)/N_{\text{total}}}$.

4. A Synthetic Example

We now instantiate the general framework from Section 3 using synthetic data. While our framework has been successfully applied to real business datasets (see Section 7), we present results using synthetic data for reproducibility and to demonstrate the approach under controlled conditions. The synthetic dataset is designed to reflect realistic business scenarios, incorporating the types of temporal patterns and cohort dynamics commonly observed in practice.

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

4.1. **Exploratory Data Analysis.** Before fitting our model, we conduct exploratory data analysis to understand the key patterns in the data. This analysis validates our modeling choices and provides a baseline for evaluating model performance. Figure 1 (shown in the Introduction) 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 7.

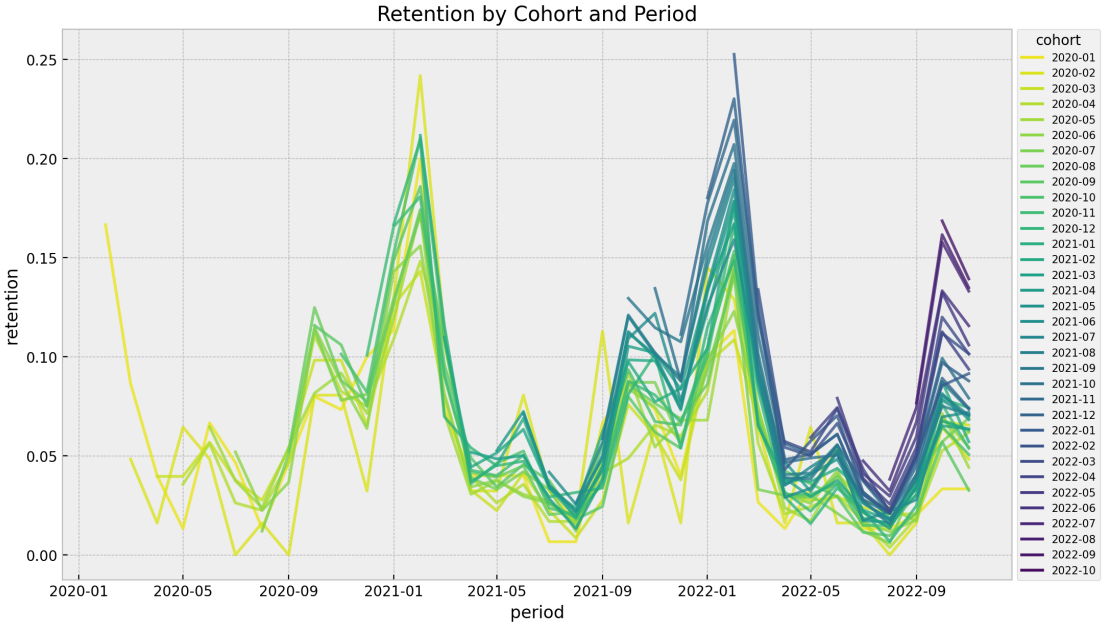

FIGURE 7. Retention as a function of the period, demonstrating the yearly seasonality pattern in retention values.

(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 the use of the number of active users (rather than the raw rate) in our likelihood, as the Binomial distribution naturally accounts for this sample-size-dependent uncertainty. As shown in Figure 3 (in the Introduction), 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 remains relatively stable over time. To verify this, we compute revenue per user as a function of age and period (Figure 8) as well as revenue per *active* user (Figure 9). 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. 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.

[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.

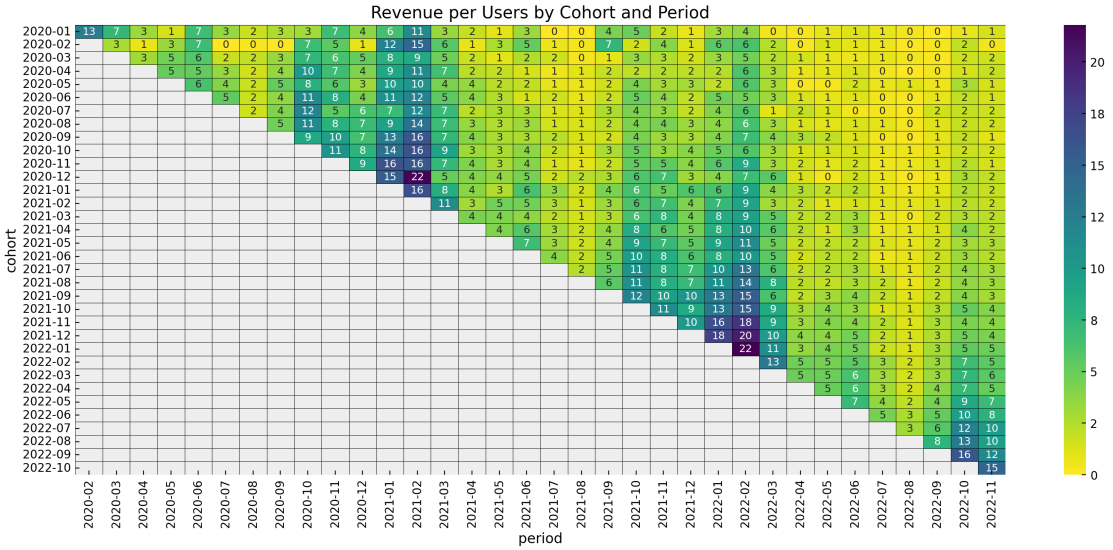

FIGURE 8. Revenue per user across cohorts. This visualization normalizes the total revenue by the original cohort size, showing the average revenue generated per initially acquired user.

- 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 validate our modeling strategy. The strong seasonality in retention motivates the inclusion of month (period) as a feature in the retention 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.

4.2. **Model Instantiation.** For this synthetic example, we instantiate the general framework using BART for the retention component and a linear model for the revenue component. This combination represents our recommended baseline.

4.2.1. *Retention Component with BART.* We model the retention probability using BART:

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

The BART component models the unknown retention latent variable 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

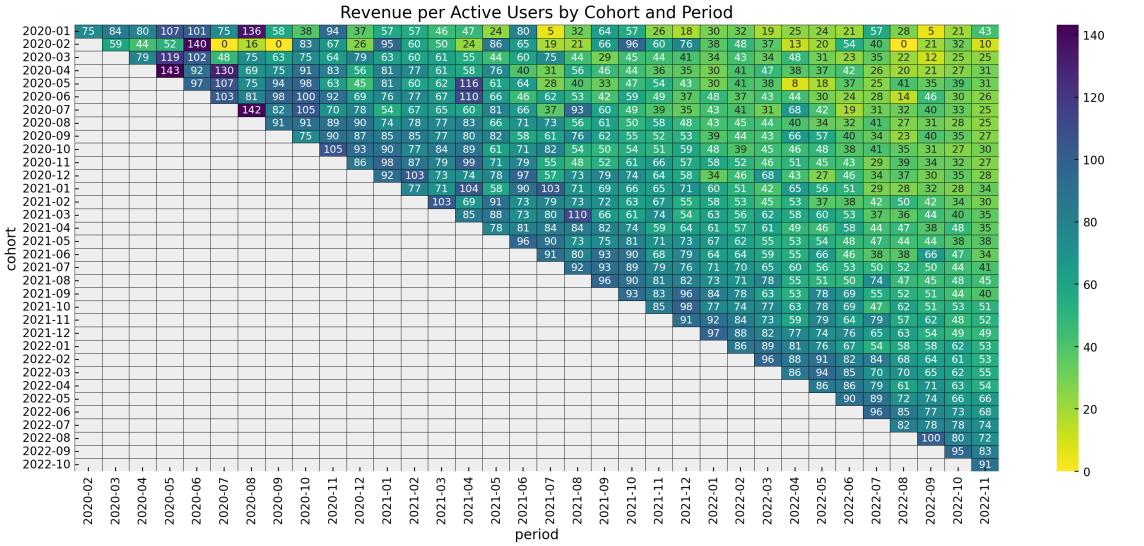

FIGURE 9. Revenue per active user across cohorts. This metric divides total revenue by the number of active users in each period, isolating spending patterns from retention effects.

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. 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 2 (Additional Covariates)**.** 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 3 (Baseline Linear Model)**.** We always recommend starting with a simple baseline, such as a linear model

$$\begin{aligned}\text{logit}(p) = &(\text{intercept} + \beta_{\text{cohort age}} \times \text{cohort age} + \beta_{\text{age}} \times \text{age} \\ &+\beta_{\text{cohort age}\times\text{age}} \times \text{cohort age} \times \text{age} + \beta_{\text{month}} \times \text{month})\end{aligned}$$

as described in [Orduz, 2023b]. In many cases, these baseline models are sufficient for the first iteration to get the decision-making process moving forward. Nevertheless, in many real applications, we have seen that the BART models can significantly improve the performance of the model (in and out of sample predictions).

4.2.2. *Revenue Component with Linear Model.* 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.

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]).

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

Remark 4 (Cohort Age Encoding)**.** 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.

4.2.3. *Complete Model Specification with Priors.* In summary, for this synthetic example, the cohort-revenue-retention model is specified as:

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

4.3. **Model Fitting and 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], where we follow the *Bayesian workflow* ([Gelman et al., 2020]). After fitting the model using MCMC, we conduct thorough diagnostics to ensure reliable inference.

Figure 10 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 (red) 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 11 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.

4.4. **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

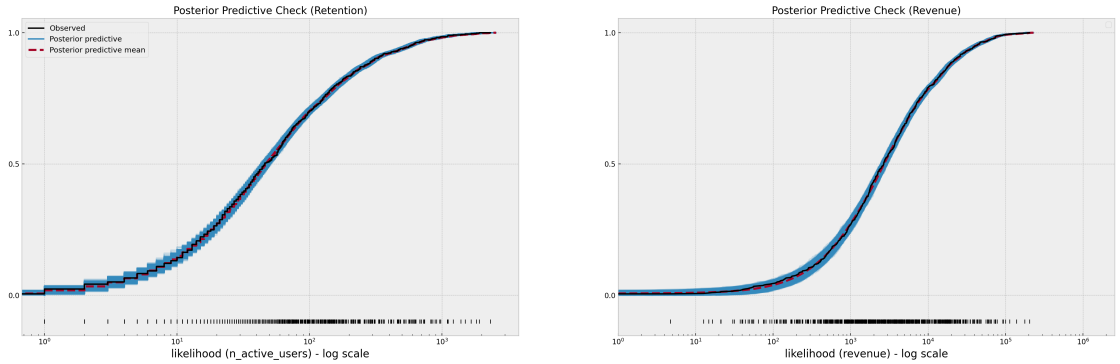

FIGURE 10. Posterior predictive distribution of the retention (left) and revenue (right) components, showing good fit to the observed data. These cumulative density plots compare the distributions of observed values (black) with simulated values from the posterior predictive distribution (orange), providing a visual assessment of model fit.

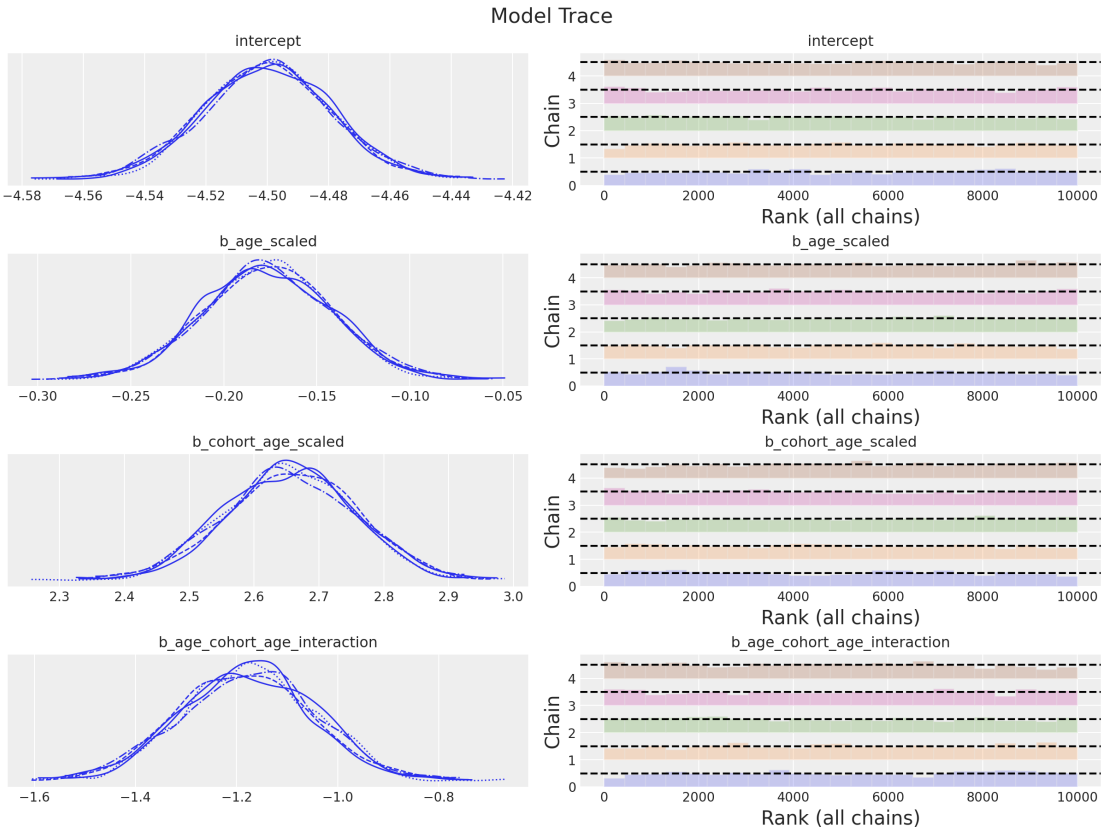

FIGURE 11. Trace plots for the linear model parameters, showing good mixing and convergence of the MCMC chains.

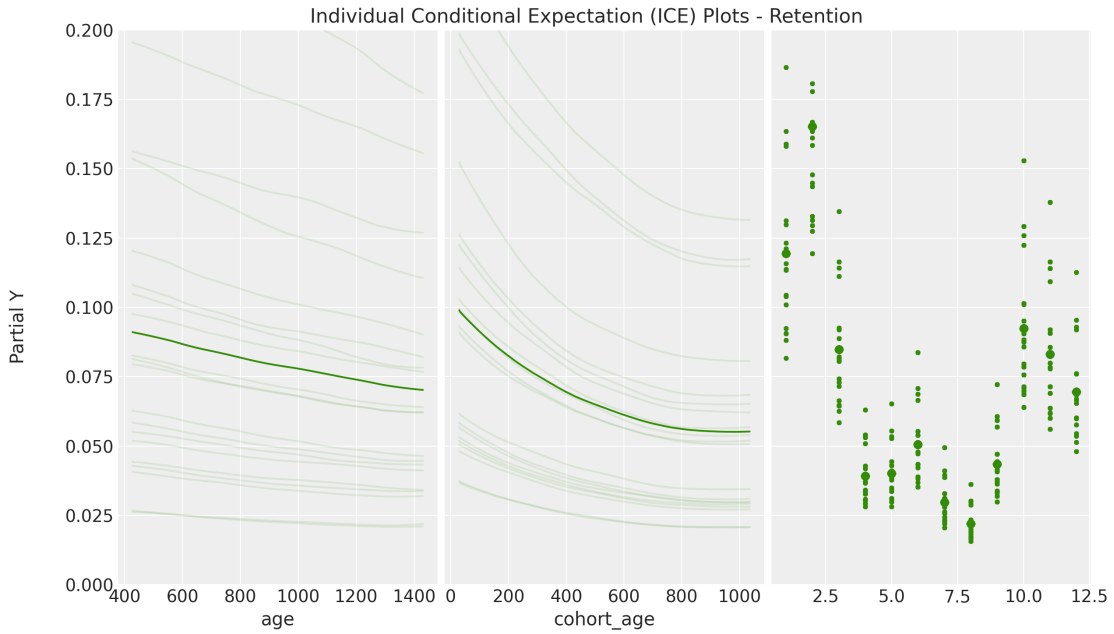

FIGURE 12. PDP (solid line) and ICE (dashed lines) plots for the retention component.

and extract actionable insights about retention drivers. Figure 12 presents *Partial Dependence Plot (PDP)* and *Individual Conditional Expectation (ICE)* 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 solid line represents the PDP plot, which is the average of the ICE plots. 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 exploratory data analysis.
- 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.

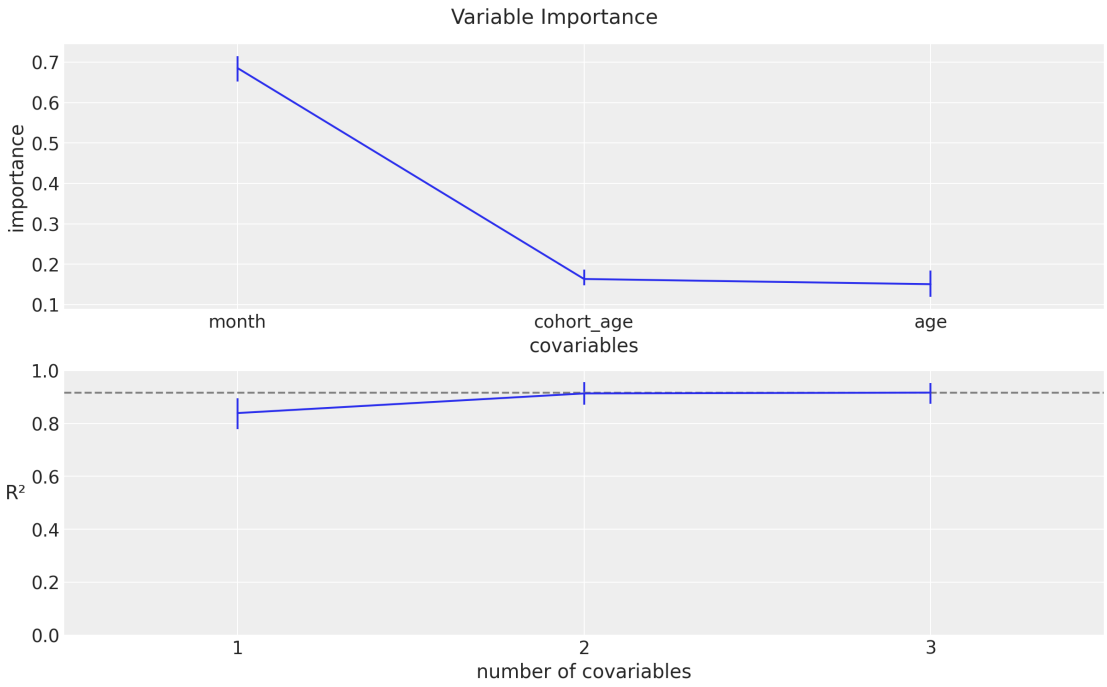

FIGURE 13. Variable importance for the retention component based on the in-sample R^2.

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 13.

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

4.5. **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.

4.5.1. *In-Sample Predictions.* We first evaluate the model's in-sample performance by comparing the posterior predictive mean against the observed values. Figure 14 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 15 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 16 shows the posterior predictive distribution compared to actual revenue values. The model successfully captures the revenue variability across different cohorts and time

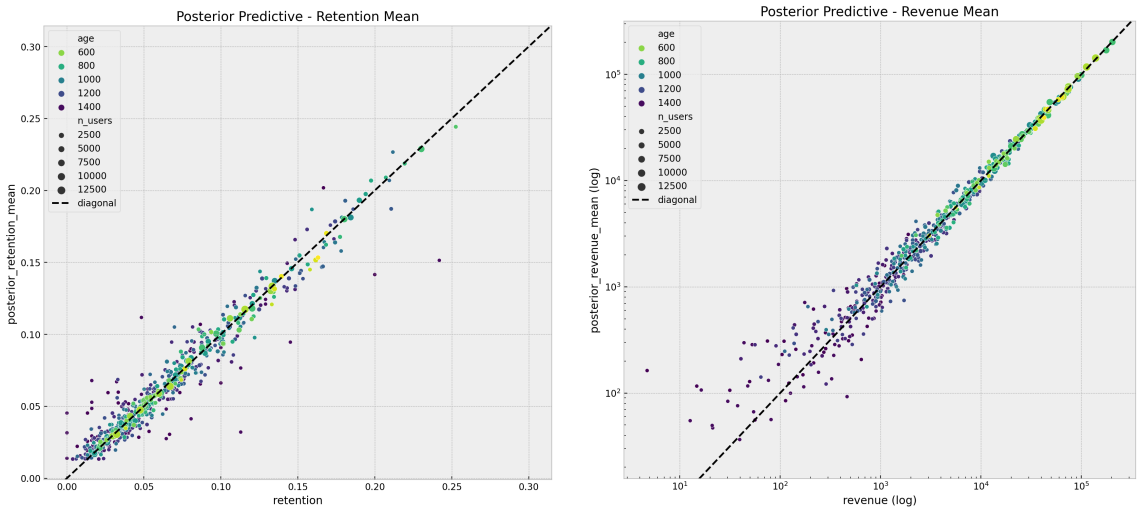

FIGURE 14. Retention (left) and revenue (right) in-sample posterior predictive mean values plotted against the actual observations. These scatter plots provide a quantitative assessment of model fit by comparing predicted versus observed values, with points closer to the diagonal line indicating better predictions.

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 17.

4.5.2. *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 18 and 19 show the out-of-sample predictions for retention and revenue, respectively. 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.
(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

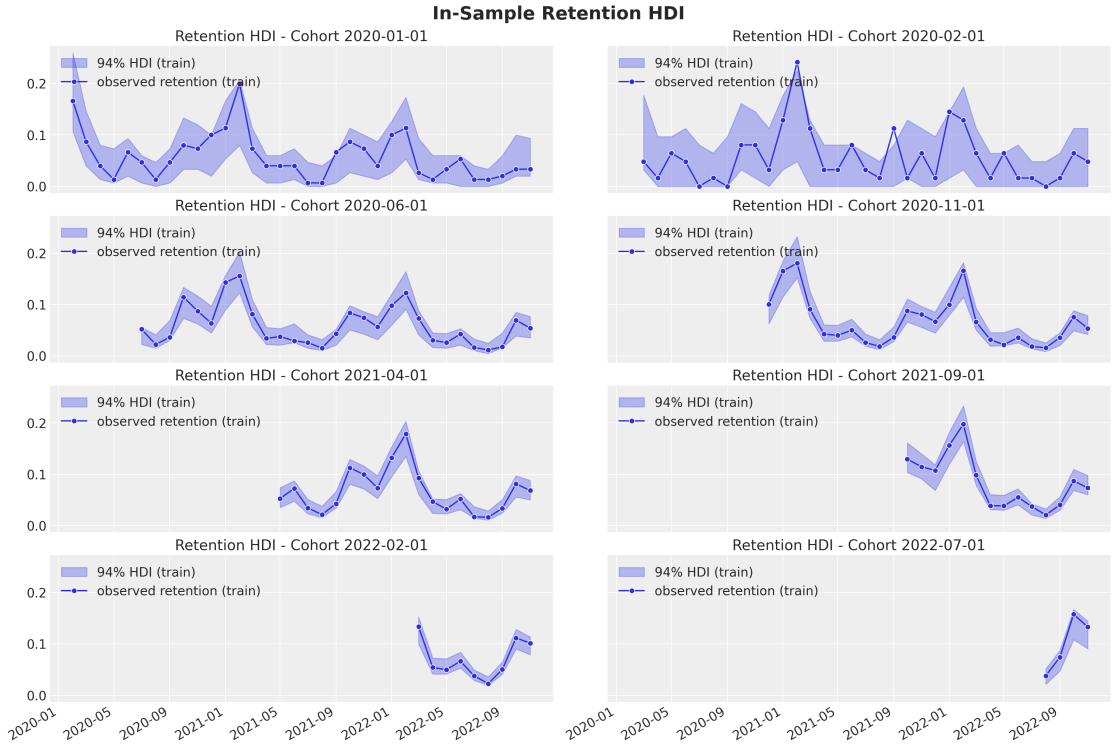

FIGURE 15. Retention in-sample posterior predictive distribution for selected cohorts, showing 94% HDI (blue shaded areas) and observed retention values (blue points). This visualization displays the model's predictive performance for retention across time for different cohorts, with uncertainty quantified 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.

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.

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

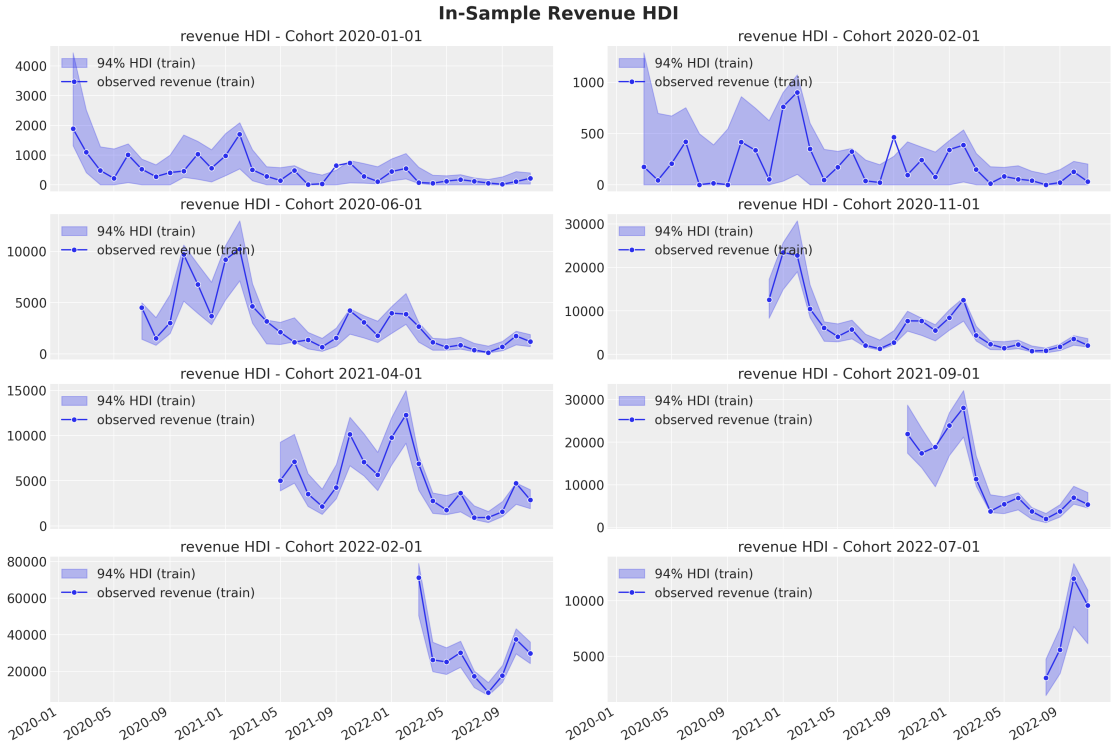

FIGURE 16. Revenue in-sample posterior predictive distribution for selected cohorts, showing 94% HDI (blue shaded areas) and observed revenue values (blue points). These plots illustrate the model's revenue predictions and associated uncertainty across time for different cohorts. The successful capture 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.

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.

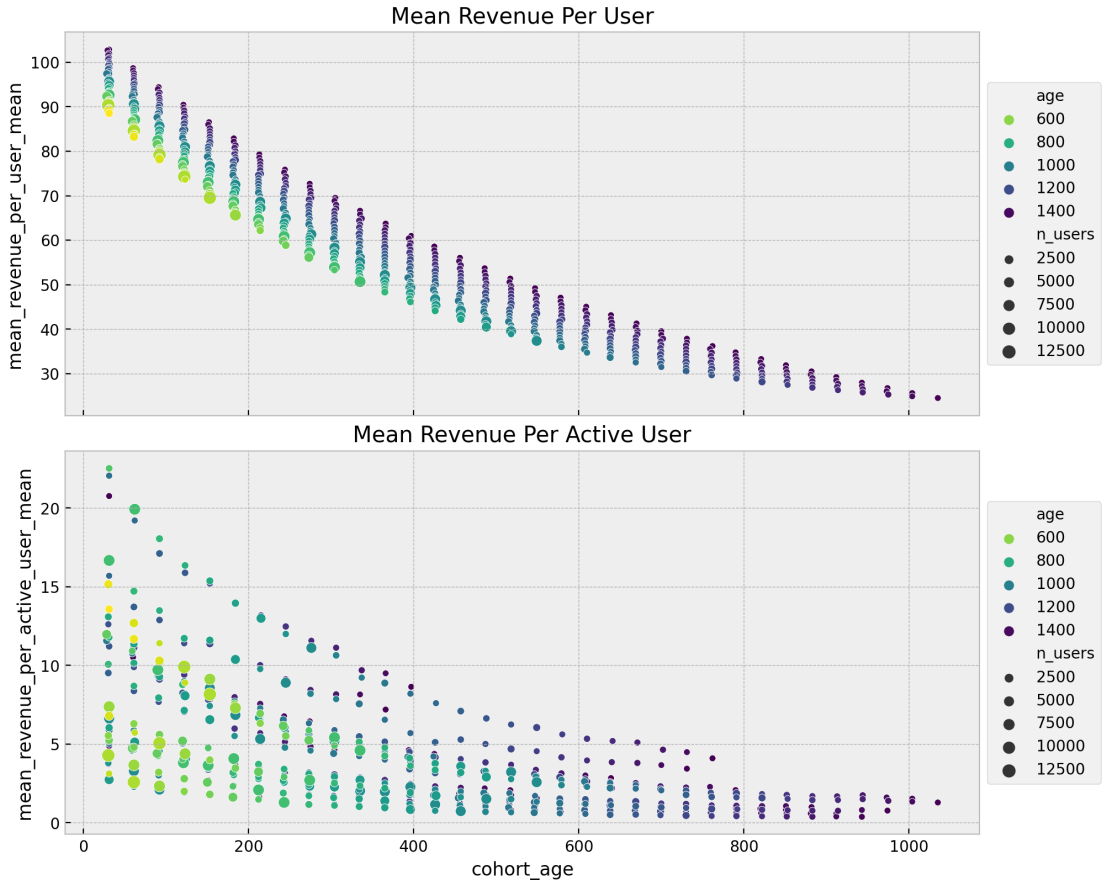

FIGURE 17. Additional view of posterior predictions across cohorts, illustrating the model's ability to capture cohort-specific patterns. This panel view organizes predictions by cohort (columns) and shows how the model adapts to the unique characteristics of each customer group.

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 ([Heek et al., 2024]), with inference performed using NumPyro ([Phan et al., 2019]). The modified model structure becomes:

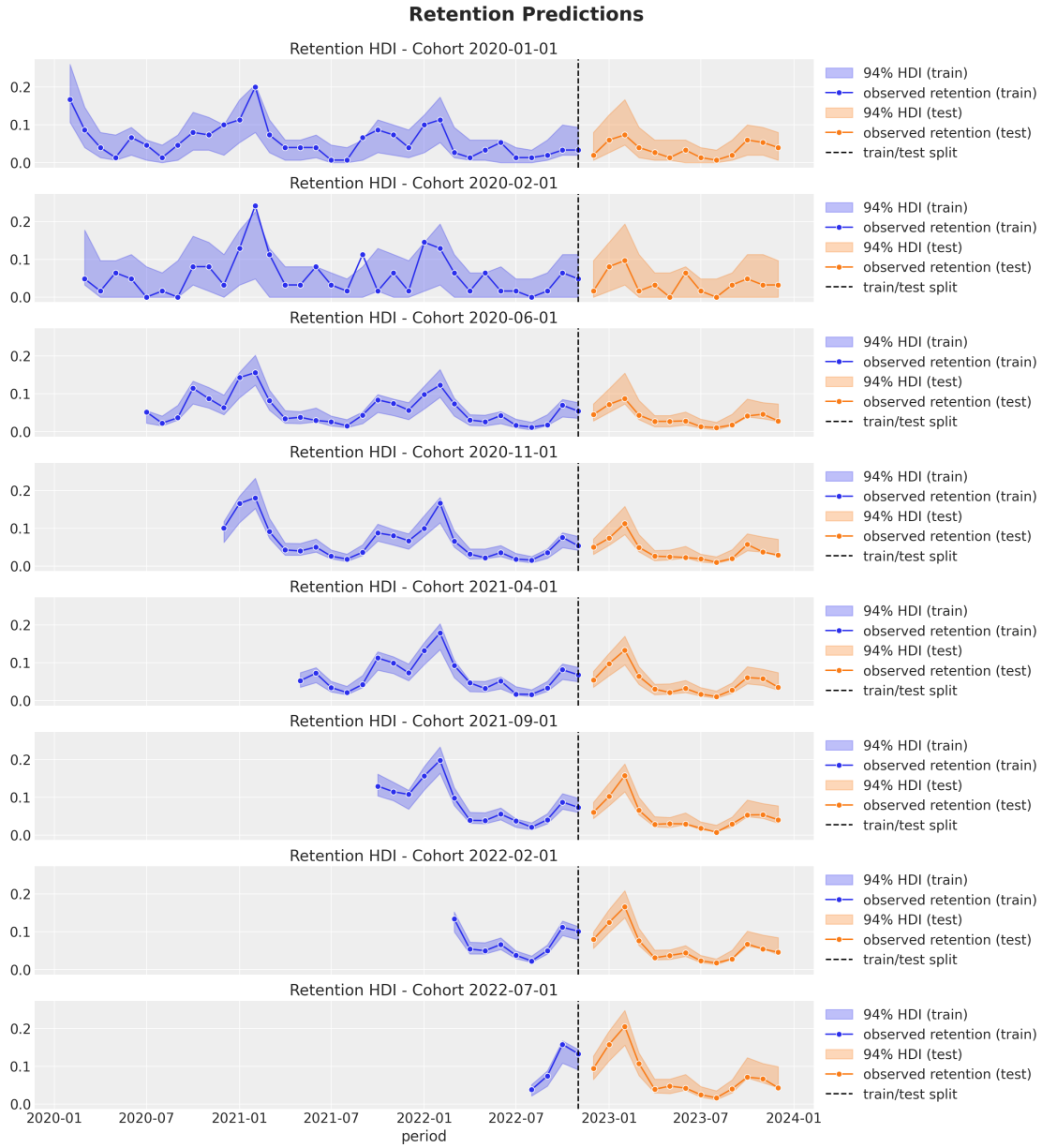

FIGURE 18. Retention out-of-sample posterior predictive distribution for (random) selected cohorts.

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

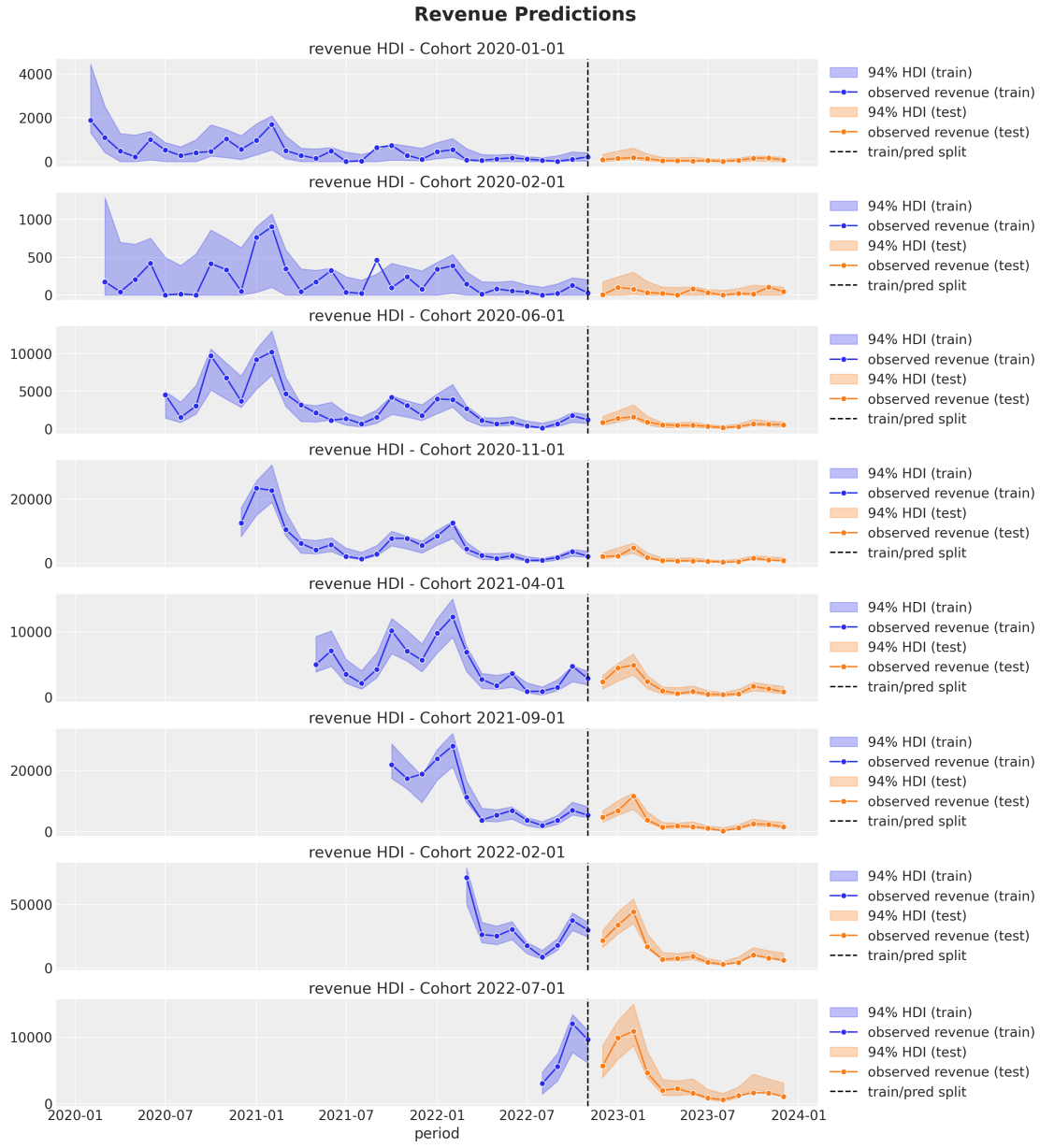

FIGURE 19. Revenue out-of-sample posterior predictive distribution for (random) selected cohorts.

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)[4]:

- **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)

[4]The code to generate the synthetic data is also available in [Orduz, 2023c] to ensure reproducibility.

The data generation process maintains the same retention and revenue dynamics described in Section 4, 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 20 visualizes the data structure, showing cohort availability and revenue patterns across the four markets.

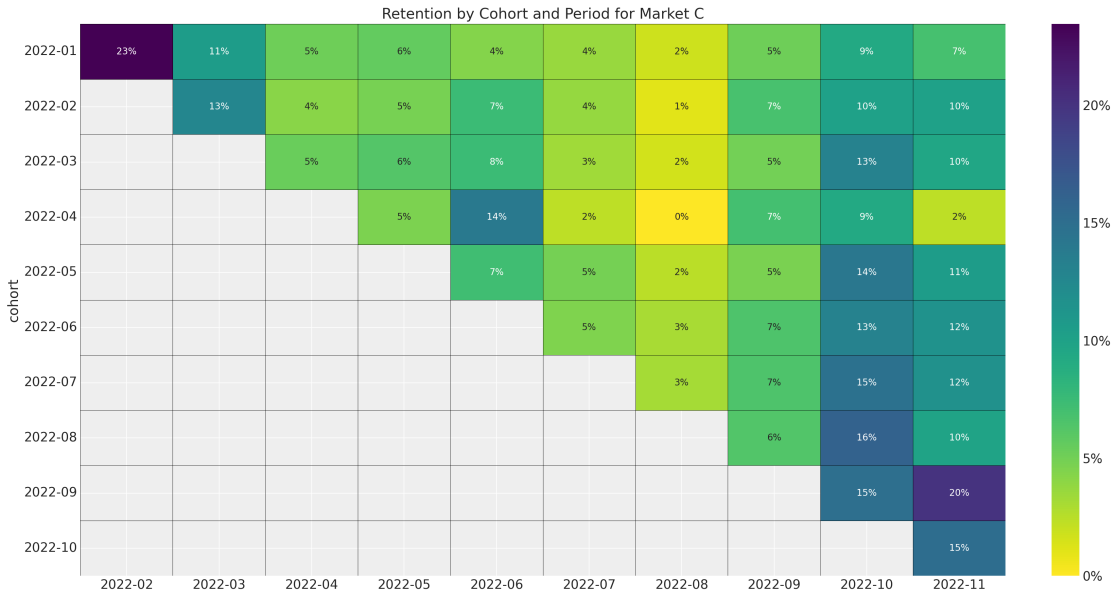

FIGURE 20. Retention matrix for Market C. We just have 10 cohorts of data available for this market.

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 21 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 and less of a year of data.

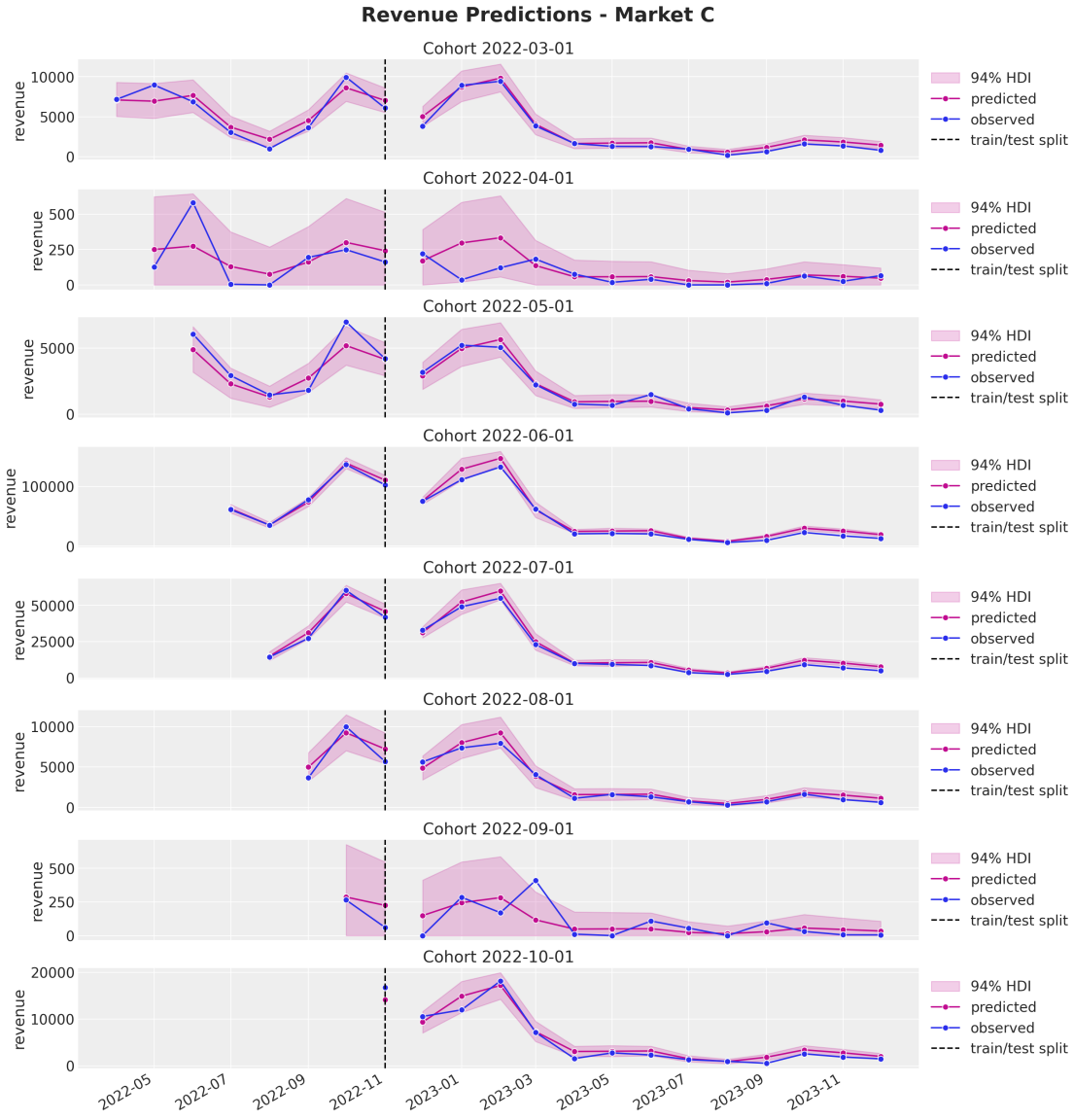

FIGURE 21. Revenue predictions for four selected cohorts in Market C.

Several key findings emerge from the hierarchical implementation:

- **Successful information pooling**: Market C achieves prediction accuracy that would be impossible with independent modeling. Note that the forecast contains the seasonal component, which would be impossible to capture with a single model on Market C as we do not even have a complete year of data. 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 C than for more established markets (e.g. Market A, for which the results look almost the same as in Section 4), 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 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[5]. For applications with fewer markets (say, $5 - 10$), MCMC remains a viable alternative that may provide more accurate uncertainty quantification. 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. That being said, inference on hierarchical models is in general more challenging becuase of the non-linear geometry of the parameter space. Fortunately, NumPyro offers automatic mechanism to reparameterize the model to make it more amenable to SVI ([Gorinova et al., 2020]).

7. A Real-Dataset Application: The H&M Transactions Dataset

Having demonstrated our framework on synthetic data with known ground truth, we now validate its practical applicability using a large-scale real-world dataset. We apply our cohort-revenue-retention model to the H&M Personalized Fashion Recommendations dataset [H&M Group, 2022], a publicly available dataset from a Kaggle competition containing transaction records from the H&M retail chain. This dataset provides an excellent testbed for our methodology due to its scale, temporal coverage, and the presence of realistic patterns in customer purchasing behavior. The dataset contains approximately 31 million transactions from over 1.36 million unique customers, spanning from September

[5]SVI is knows to underestimate the uncertainty, but it is still a valid approximation for the posterior distribution.

2018 to September 2020. For our cohort analysis, we aggregate customers into monthly cohorts based on their first purchase date, resulting in 20 cohorts from November 2018 to June 2020. We use a train/test split at May 2020, holding out 3 months of data (June–August 2020) for out-of-sample evaluation. This split allows us to assess the model's forecasting ability while ensuring sufficient training data for each cohort. Figures 22 and 23 present the retention and revenue matrices for the H&M dataset. Several patterns emerge that are characteristic of real retail data:

- Retention rates show a clear decay pattern as cohort age increases, with the steepest decline occurring in the first few months after customer acquisition.
- Revenue patterns closely track the number of active users, consistent with our modeling assumption that revenue can be decomposed into active user counts and revenue per active user.

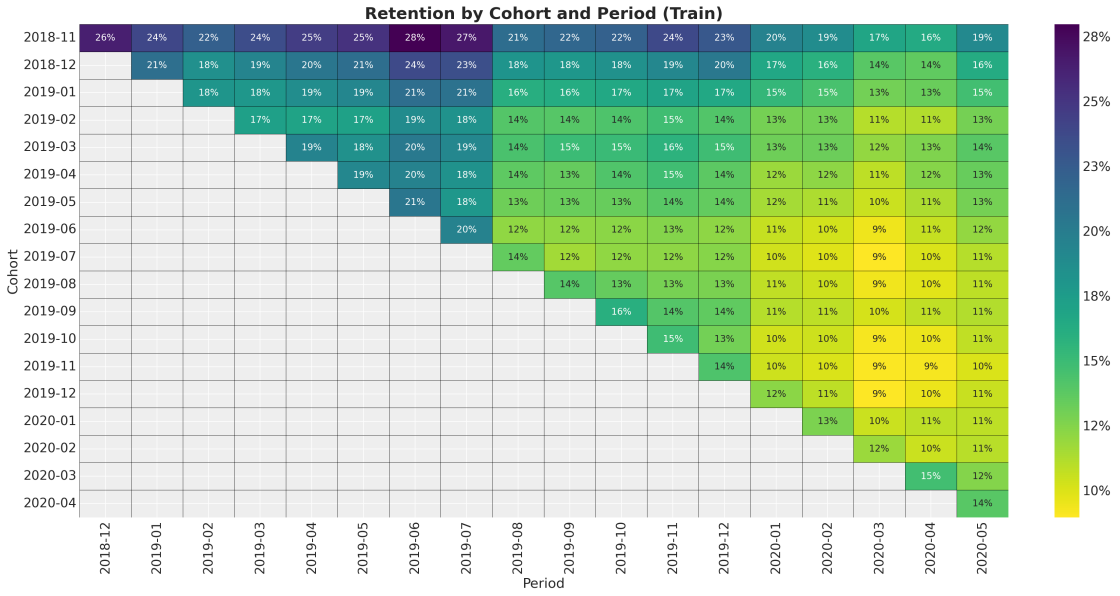

FIGURE 22. Retention matrix for the H&M dataset (training period).

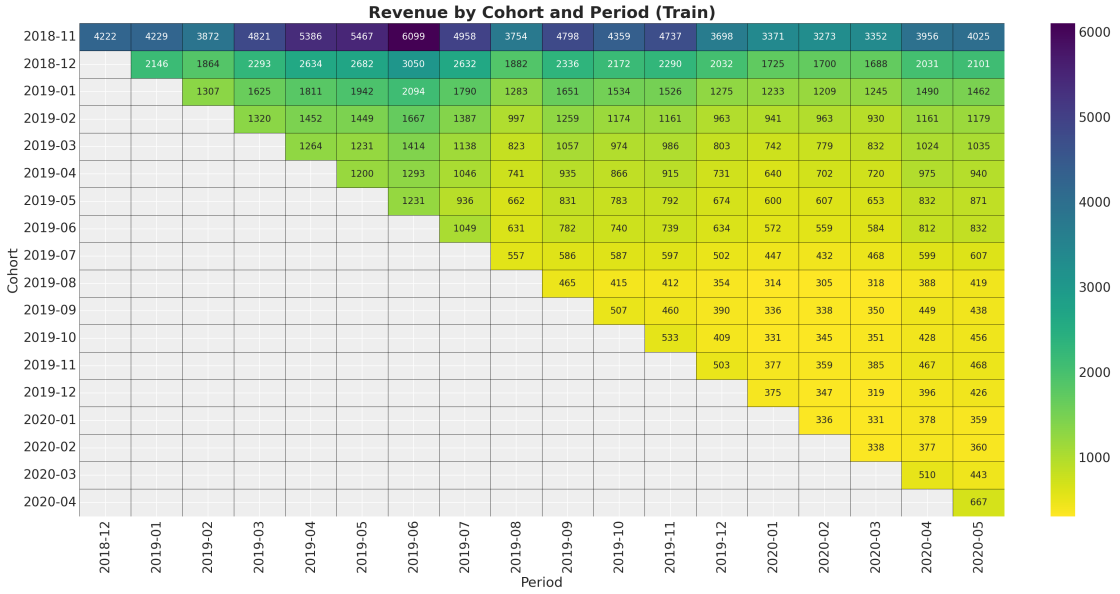

FIGURE 23. Revenue matrix for the H&M dataset (training period).

We benchmark our Bayesian cohort-revenue-retention model against three established approaches commonly used for time series forecasting in business contexts (using two established data science libraries: the Nixtla forecasting framework [Nixtla, 2023] and the `scikit-learn` library [Pedregosa et al., 2011]):[6]

(1) **AutoMFLES**: An automated multiple seasonal-trend decomposition model from the 'StatsForecast' library that automatically detects and models multiple seasonalities.
(2) **Simple Exponential Smoothing (SES)**: A classical time series method that produces forecasts as weighted averages of past observations, with weights decaying exponentially.
(3) **HistGradientBoostingRegressor**: A gradient boosting machine learning model from 'scikit-learn', representing modern ML approaches to time series forecasting.

These baselines represent the spectrum of approaches typically considered for cohort-level revenue forecasting.

[6]The notebook to reproduce the benchmark results is available at `https://github.com/juanitorduz/website_projects/blob/master/Python/cohort-revenue-retention/hm-transactions.ipynb`.

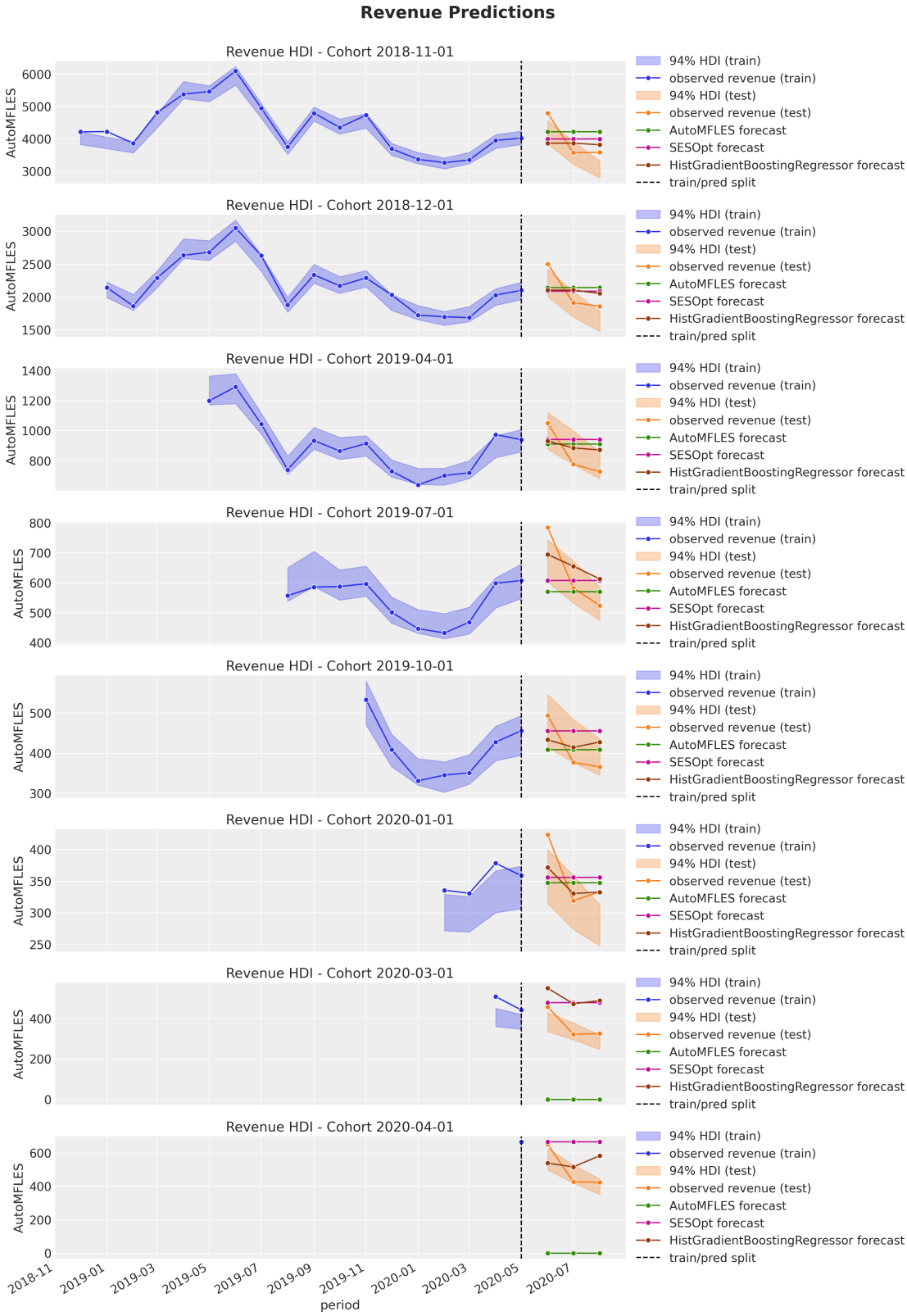

FIGURE 24. Revenue predictions comparing the Bayesian model against other benchmark approaches for a selected set of cohorts.

Figure 24 shows revenue predictions for selected cohorts for the Bayesian cohort-revenue-retention model against the other benchmark model. One can clearly see that the Bayesian model is able to capture the dynamics with well-calibrated 94% highest density intervals, while the other benchmark models fail to efficiently use the data to generate accurate predictions. The reason is that the coupling between retention and revenue provides key signals to the revenue model that are hard to get from the revenue data, even if we allow arbitrary non-linear relationships such that the gradient boosting model `HistGradientBoostingRegressor` can learn. The failure of other benchmark methods is particularly pronounced for cohorts with limited training history (e.g., the March 2020 and April 2020 cohorts, which have only 2–3 observations before the test period). These methods cannot extrapolate meaningful patterns from such sparse data. Our framework, by contrast, leverages the shared structure across all cohorts, learning that retention decays with cohort age, that revenue correlates with active users, and that seasonal patterns affect all cohorts similarly to generate reliable predictions even for data-sparse cohorts. Table 1 presents quantitative evaluation metrics comparing out-of-sample prediction accuracy across all models. We report Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) computed on the held-out test set.

TABLE 1. Out-of-sample evaluation metrics for revenue prediction on the H&M dataset. Lower values indicate better predictive performance. The Bayesian model substantially outperforms all other approaches.

Model	**MAE**	**RMSE**	**MAPE (%)**
Bayesian (BART)	82.46	135.64	10.03
AutoMFLES	190.39	256.97	28.07
SES	142.64	193.19	17.78
HistGradientBoostingRegressor	161.90	218.32	25.12

The quantitative results confirm what the visual comparisons suggest: our Bayesian cohort-revenue-retention model dramatically outperforms standard time series and machine learning approaches. The improvement is most pronounced in MAPE, where the Bayesian model achieves 10.03% compared to 17.78%–28.07% for the baselines, reflecting the model's ability to maintain proportionally accurate predictions across cohorts of varying sizes and revenue magnitudes. The MAE and RMSE metrics tell a similar story: the Bayesian model's MAE of 82.46 is roughly half that of the best-performing baseline (SES at 142.64), demonstrating substantial improvements in absolute prediction accuracy.

Finally, Figure 25 aggregates predictions across all training cohorts to show total revenue forecasting performance. The Bayesian model's predictions (with 50% and 94% HDI bands) track the actual revenue trajectory, while baseline methods diverge substantially in the test period.

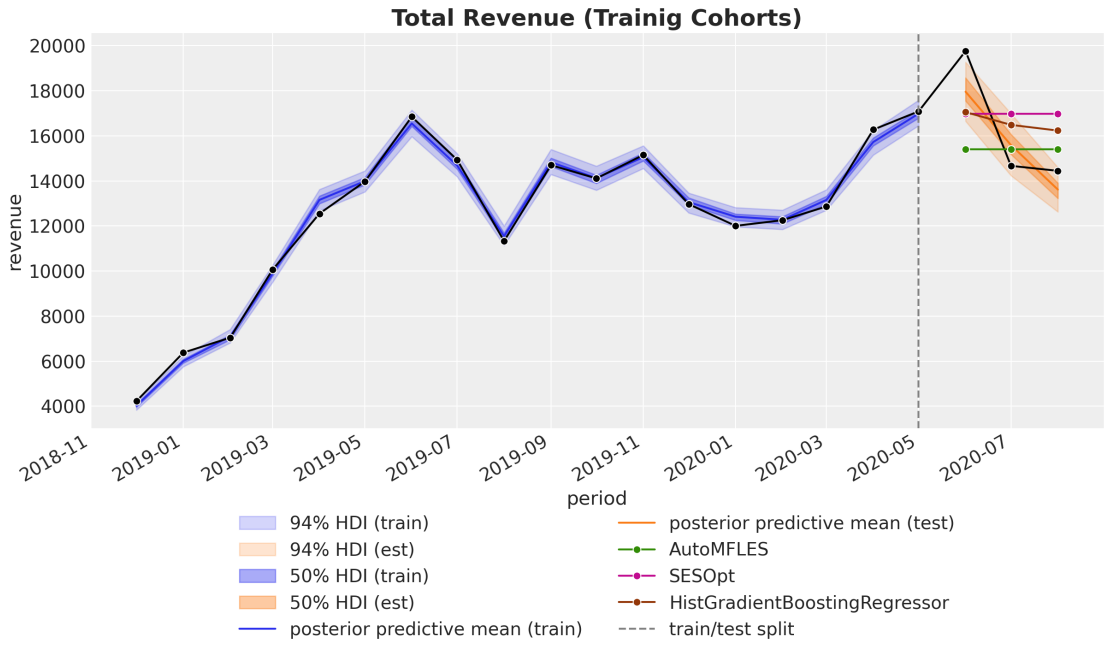

FIGURE 25. Total revenue predictions (aggregated across all training cohorts) comparing the Bayesian model with baselines. The black line shows observed total revenue, while colored lines show model predictions.

Remark 5. It is worth noting that individual-level probabilistic models, such as the Shifted Beta Geometric (sBG) model [Fader and Hardie, 2007a] implemented in PyMC-Marketing [PyMC-Labs, 2023], provide an alternative framework for modeling customer retention in contractual settings. However, fitting such models using MCMC on datasets of this scale with over 1.36 million unique customers and 31 million transactions is computationally prohibitive without substantial investment in specialized hardware infrastructure (e.g., high-end GPUs). While stochastic variational inference can improve scalability, the cohort-level approach presented in this paper offers a practical and effective alternative: by aggregating to the cohort level, we dramatically reduce the computational burden while retaining the ability to capture complex temporal patterns and provide well-calibrated uncertainty estimates. The benchmark results above demonstrate that this cohort-level framework not only scales gracefully but also delivers superior predictive performance compared to standard time series and machine learning baselines.

7.1. **Comparison with Classical CLV Models.** To provide a comprehensive evaluation of our proposed framework, we compare it against the classical buy-till-you-die (BTYD) models from the customer lifetime value literature. Specifically, we implement the BG/NBD model [Fader et al., 2005a] for modeling purchase frequency and dropout,

combined with the Gamma-Gamma model [Fader et al., 2005b] for monetary value estimation. These models represent the standard approach for CLV estimation in non-contractual settings and are implemented in PyMC-Marketing [PyMC-Labs, 2023][7].

7.1.1. *Cohort-Based Fitting Approach.* A key challenge when applying BG/NBD models to retail datasets is their assumption of stationary purchase rates. The model assumes that each customer's underlying transaction rate remains constant over time, which conflicts with the strong seasonal patterns observed in fashion retail data. The H&M dataset exhibits clear seasonality in both retention and revenue (as shown in Figures 22 and 23), making a single global BG/NBD model inappropriate.

To address this limitation, we adopt a cohort-based fitting strategy: instead of fitting a single model to all 1.36 million customers, we fit separate BG/NBD and Gamma-Gamma model pairs for each of the 20 monthly cohorts. This approach implicitly captures some seasonal effects by allowing each cohort to have its own parameter estimates, reflecting the purchasing patterns of customers acquired during different times of the year. We use maximum a posteriori (MAP) estimation because full MCMC sampling at this scale, with over one million customers, is computationally infeasible in practice. Moreover, given the model's inherent limitations (stationarity assumption, lack of seasonality flexibility), investing substantial computational resources in full Bayesian inference would not address the fundamental modeling constraints.

7.1.2. *Implementation Challenges and Solutions.* Fitting BG/NBD models via MAP estimation proved challenging for several cohorts with extreme data characteristics. The optimizer can produce NaN parameter estimates when cohorts exhibit:

- **High repeat rates** ($> 90\%$): The optimizer pushes the dropout parameters (a and b in the Beta distribution governing dropout probability) toward extreme values, causing numerical instability.
- **Low recency/T ratios** (< 0.25): Customers who made their last purchase long before the observation end appear "dormant," leading to extreme estimates for the dropout shape parameter.

To address these convergence issues, we implemented a progressive prior tightening strategy with three tiers of increasingly informative priors:

(1) **Default priors** ($\sigma = 5$): Moderately informative HalfNormal priors that work for typical cohorts with balanced repeat rates.
(2) **Tight priors** ($\sigma = 2$): Used as a fallback for cohorts with high-frequency customers (mean frequency > 20), constraining the parameter space to avoid extreme values.
(3) **Very tight priors** ($\sigma = 1$): A final fallback for cohorts with low recency/T ratios, strongly constraining parameters near typical values.

The fitting algorithm attempts each configuration in sequence, with multiple random seeds per configuration, until a valid (non-NaN) fit is obtained. This strategy successfully

[7]The complete implementation code for this comparison is available at `https://github.com/juanitorduz/website_projects/blob/master/Python/cohort-revenue-retention/hm-pymc-marketing.ipynb`.

fitted all 20 cohorts, though we note the trade-off: tighter priors stabilize MAP estimation but may introduce bias in parameter estimates.

7.1.3. *Results and Discussion.* Figure 26 presents the revenue predictions from the cohort-based BG/NBD + Gamma-Gamma models for the same cohorts shown in Figure 24. Several qualitative observations emerge from comparing this approach with our BART-based framework:

FIGURE 26. Revenue predictions using cohort-based BG/NBD + Gamma-Gamma models for selected cohorts. Blue lines show observed training revenue, orange lines show model predictions, and orange dots indicate observed test revenue.

- **Flat predictions**: The BG/NBD predictions are essentially flat across the test period due to the stationarity assumption, unable to capture seasonal patterns visible in the actual data.
- **No cross-cohort information sharing**: Each cohort's model is fitted independently, which is particularly problematic for young cohorts with limited training data.

Classical BTYD models remain valuable tools for individual-level CLV estimation in stationary settings. However, for cohort-level and aggregate revenue forecasting, the focus of this paper, our framework offers practical advantages: it is more accurate for seasonal data, simpler to implement (a single model versus 20 separate model pairs), and more interpretable through tools like partial dependence plots. The ability to naturally incorporate calendar effects and share information across cohorts makes it well-suited for the business decision-making contexts we target. We could, of course, think about a fully Bayesian hierarchical model where we pool in the parameters of the BG/NBD and Gamma-Gamma models across cohorts, but this would be computationally prohibitive for the scale of the dataset (which is not uncommon for many businesses).

7.2. **Limitations and Future Work.** While the proposed framework offers significant advantages in terms of flexibility and predictive accuracy, it is important to acknowledge its limitations and potential areas for future development.

First, by operating at the cohort level, the model necessarily sacrifices individual-level granularity. While this aggregation reduces noise and improves computational efficiency, it also means that specific customer-level behaviors such as individual product preferences or idiosyncratic purchase triggers are not explicitly captured. For applications requiring highly personalized interventions, this framework should be seen as a complement to, rather than a replacement for, individual-level models.

Second, the current coupling mechanism assumes that the average revenue per active user follows a relatively stable functional form $\lambda(x)$. While we demonstrated that this holds well in both synthetic and real-world datasets, sudden structural breaks in the market (e.g., major policy changes or black-swan events) might require more complex modeling of the revenue rate, potentially including stochastic volatility components.

Third, the framework's performance on extremely sparse cohorts remains an area for further investigation. While hierarchical information pooling significantly improves predictions for young cohorts, there is a fundamental lower bound on the amount of data required to distinguish between different latent drivers. Future work could explore incorporating even stronger priors or external macroeconomic indicators to further regularize these data-sparse regimes.

Finally, we have primarily focused on Gamma and Binomial likelihoods. While these are appropriate for many business contexts, exploring other distributional assumptions (e.g., Zero-Inflated models for markets with many non-purchasers) could further extend the framework's reach.

8. Conclusion

In this paper, we have presented a Bayesian framework for jointly modeling cohort-level retention and revenue. The core contribution is the coupling mechanism:

$$R_{i,j} \sim \text{Gamma}(N_{\text{active},i,j}, \lambda_{i,j})$$
$$N_{\text{active},i,j} \sim \text{Binomial}(N_{\text{total},i}, p_{i,j})$$

where the number of active users $N_{\text{active},i,j}$ estimated through the retention model directly informs the revenue model as its shape parameter. This coupling provides consistent uncertainty propagation, natural variance scaling, and an interpretable mean structure where $1/\lambda_{i,j}$ represents the average revenue per active user.

A key strength of this framework is its flexibility: the latent variables p (retention probability) and λ (revenue rate) can be modeled using various approaches. We demonstrated BART as a recommended baseline for the retention component due to its interpretability through PDP/ICE plots, but also showed that neural networks offer a computationally efficient alternative for large-scale applications. The framework naturally extends to hierarchical multi-market settings, enabling information pooling across markets with varying data availability.

As discussed in our review of related work (Section 2), 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 the coupled retention-revenue architecture.

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. The validation on the H&M dataset (Section 7) demonstrates that this cohort-level framework delivers superior predictive performance compared to standard time series and machine learning baselines, particularly for cohorts with limited training history where the coupling mechanism provides crucial information transfer.

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

[Fader et al., 2005b] Fader, P., Hardie, B., and Lee, K. (2005b). Rfm and clv: Using iso-value curves for customer base analysis. *Journal of Marketing Research American Marketing Association ISSN*, XLII:415–430.

[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.

[Gelman et al., 2020] Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow.

[Gorinova et al., 2020] Gorinova, M. I., Moore, D., and Hoffman, M. D. (2020). Automatic reparameterisation of probabilistic programs. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

[Heek et al., 2024] Heek, J., Levskaya, A., Oliver, A., Ritter, M., Rondepierre, B., Steiner, A., and van Zee, M. (2024). Flax: A neural network library and ecosystem for JAX.

[H&M Group, 2022] H&M Group (2022). H&M Personalized Fashion Recommendations. https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations. Kaggle Competition Dataset.

[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.

[Nixtla, 2023] Nixtla (2023). Nixtla: State-of-the-art time series forecasting. https://github.com/nixtla. Open source time series forecasting tools and libraries.

[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/.

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

[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

Email address: daniel.guhl@hu-berlin.de

Humboldt-Universität zu Berlin, Germany