COHORT REVENUE & RETENTION ANALYSIS: A BAYESIAN APPROACH

JUAN CAMILO ORDUZ

Abstract. We present a Bayesian approach to model cohort-level retention rates and revenue over time. We use Bayesian additive regression trees (BART) to model the retention component which we couple with a linear model for the revenue component. This method is flexible enough to allow adding additional covariates to both model components. This Bayesian framework allows us to quantify uncertainty in the estimation, understand the effect of covariates on retention through partial dependence plots (PDP) and individual conditional expectation (ICE) plots, and most importantly, forecast future revenue and retention rates with well-calibrated uncertainty through highest density intervals. We also provide alternative approaches to model the retention component using neural networks and inference through stochastic variational inference.

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 [4], where subscription-based relationships predominate, and non-contractual settings [3], 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 ([1]), where the PyMC-Marketing library [12] 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 just need to understand behaviors at the cohort level—groups 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:

Date: April 22, 2025.

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

(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 [6], 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.

This work introduces a Bayesian approach that addresses these challenges by modeling cohort-level retention rates from a top-down perspective. Instead of building up from individual purchase patterns—a process that can become computationally intensive and complex for large customer bases—we directly model aggregate retention and revenue at the cohort level. 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 so the same for the revenue matrix and 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:

[2]Although, one can add regressors in some cases as described in [5] for the non-contractual case.

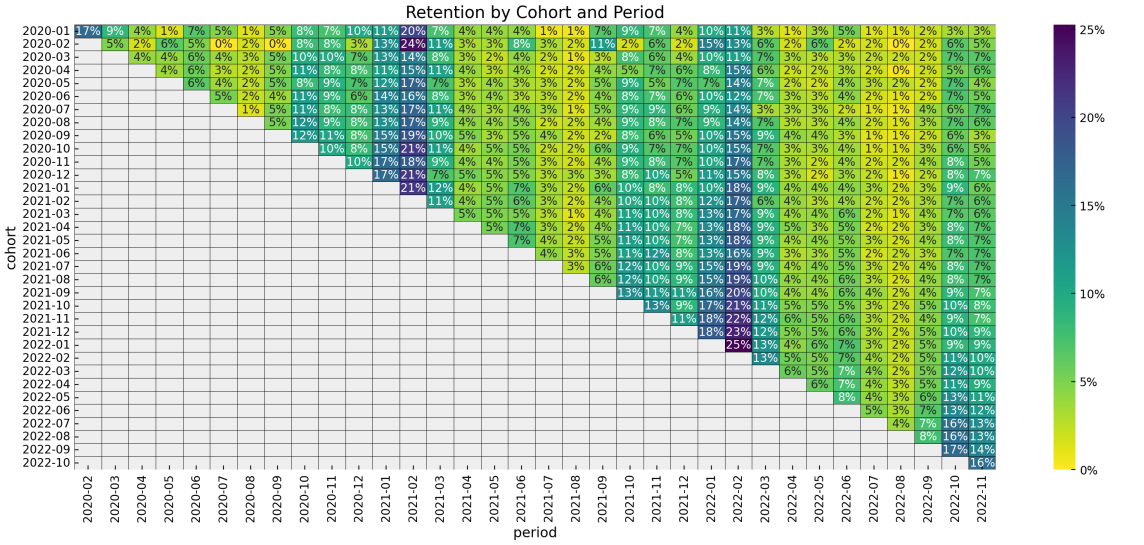

FIGURE 1. Retention matrix example. The matrix visualizes customer retention rates across different cohorts (rows) and observation periods (columns). Each cell represents the proportion of customers from a specific acquisition cohort that remained active in a subsequent period. Colors indicate retention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

- **Flexibility in relationship modeling**: By employing Bayesian additive regression trees (BART) [13], 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.
- **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.

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

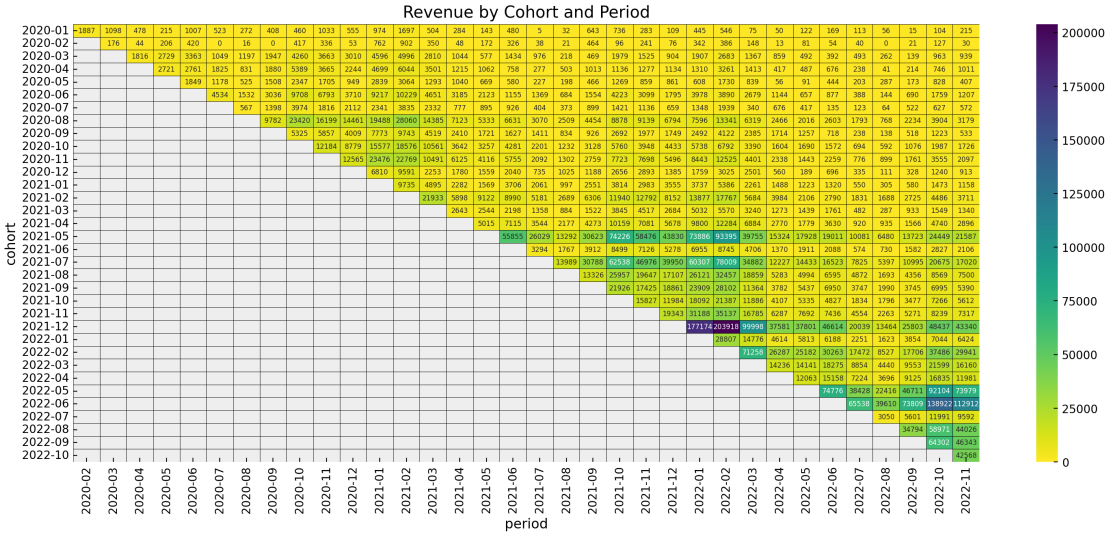

FIGURE 2. Revenue per cohort. This heatmap visualizes the total revenue generated by each cohort (rows) across different time periods (columns). The color intensity corresponds to revenue magnitude, revealing a strong correlation with the number of active users (Figure 3).

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 the most interesting (and novel) part of this work. We couple the two components through the number of active users. Here is the full model specification:

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

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.

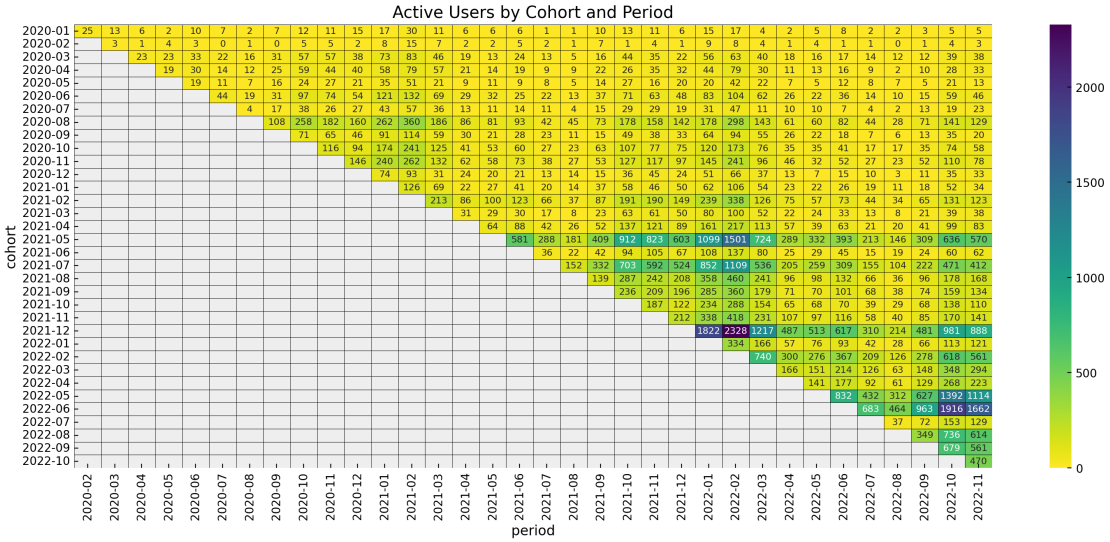

FIGURE 3. Number of active users across cohorts. This heatmap displays the absolute count of active users for each cohort (rows) across observation periods (columns).

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

To illustrate our approach, we use a synthetic dataset (available as a *csv* file from [8]). The code to generate this dataset (deterministically) is publicly available in [9]. Let's begin with exploratory data analysis. 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

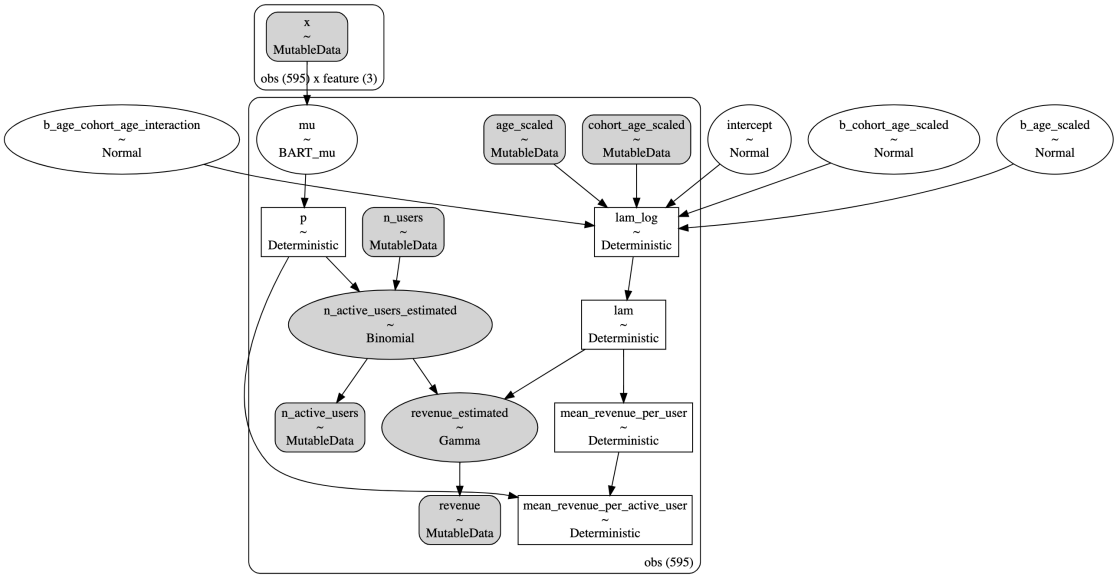

FIGURE 4. Cohort-revenue-retention model structure. This diagram illustrates the two coupled components of our model: the retention component (left) using BART to model the probability of customers remaining active, and the revenue component (right) using a gamma distribution with parameters informed by the retention model. The arrows show the flow of information, demonstrating how the estimated number of active users from the retention model directly feeds into the revenue model.

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.

With this exploratory analysis complete, we can proceed to the modeling phase.

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

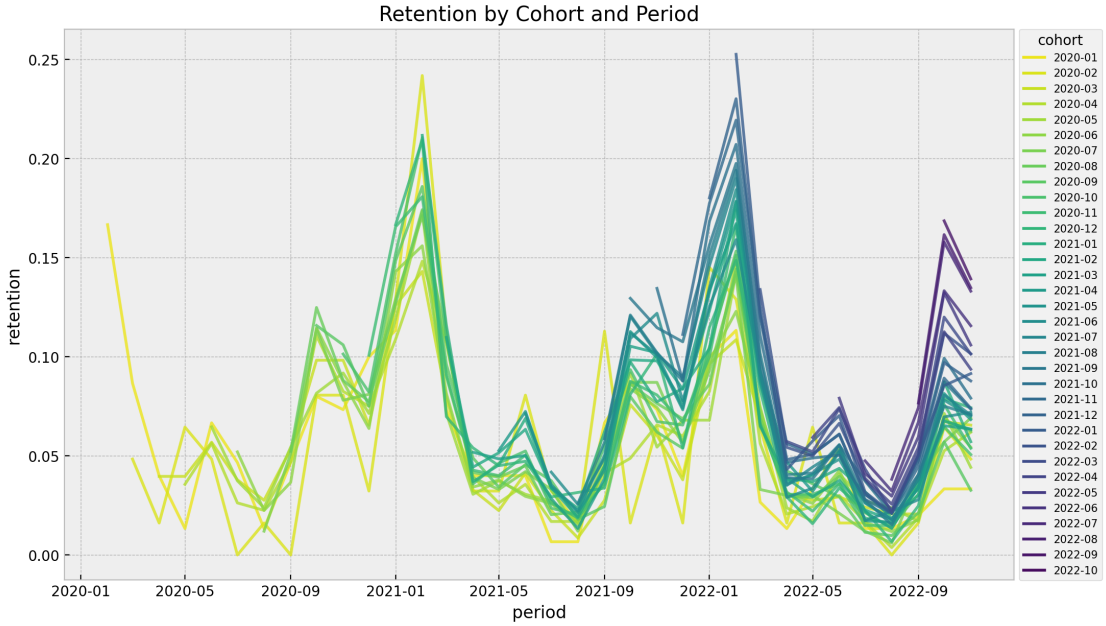

FIGURE 5. Retention as a function of the period, demonstrating the yearly seasonality pattern in retention values.

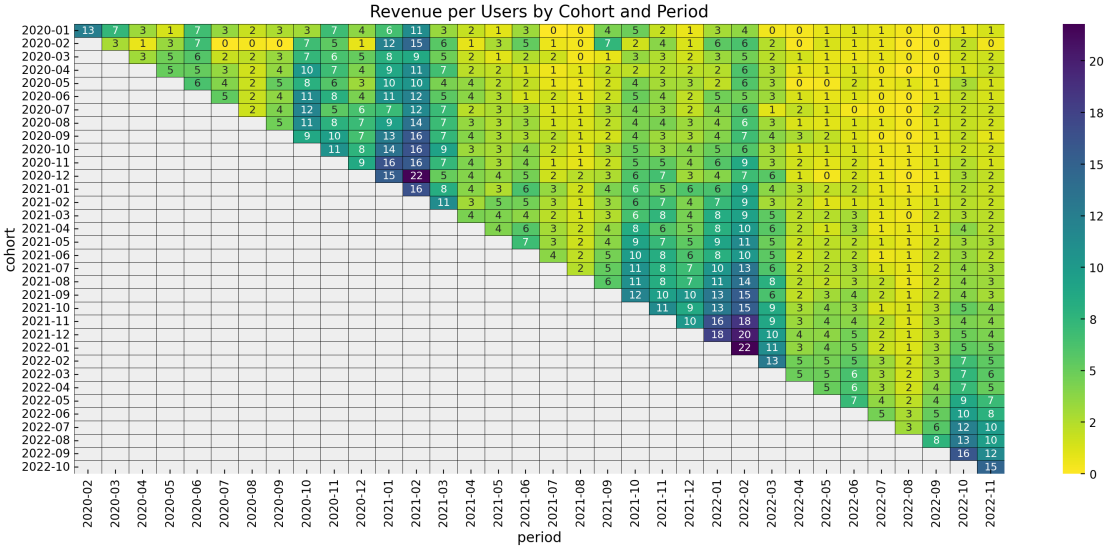

FIGURE 6. Revenue per user across cohorts. This visualization normalizes the total revenue by the original cohort size, showing the average revenue generated per initially acquired user.

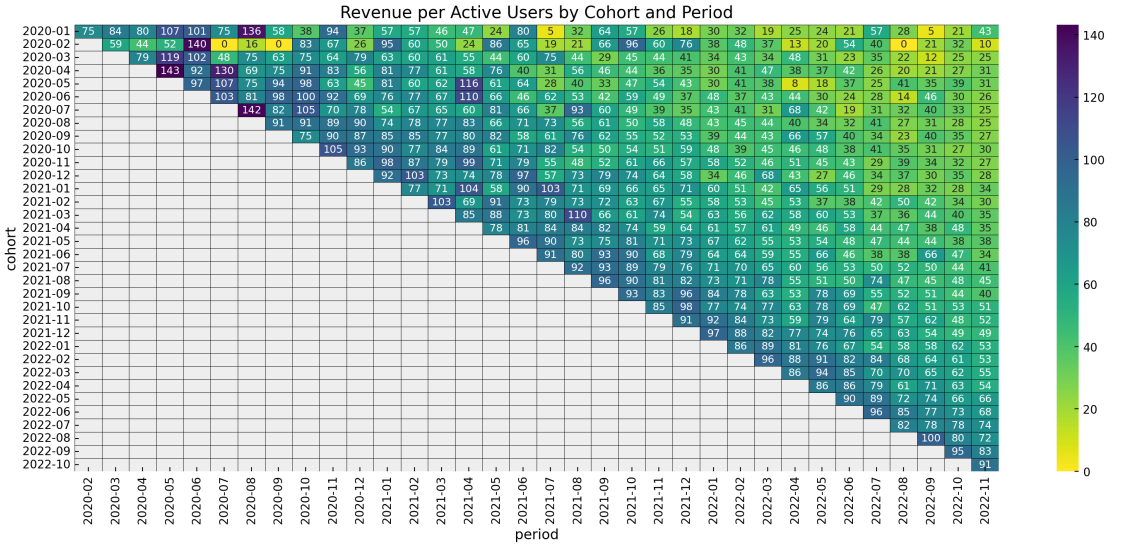

FIGURE 7. Revenue per active user across cohorts. This metric divides total revenue by the number of active users in each period, isolating spending patterns from retention effects.

3. Model Specification and Diagnostics

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 $\text{Binomial}(N_{\text{total}}, p)$, 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.

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

The main parameter we need to specify for the BART model is the number of trees. We typically start with a small number of trees and increase it incrementally while monitoring the posterior predictive distribution's quality.

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 [8], our experience with real datasets shows that such simpler approaches often fail to adequately capture the complex patterns in the data.

For the revenue component, we employ a gamma random variable $\text{Gamma}(N_{\text{active}}, \lambda)$ (inspired by [14]). The mean of this gamma distribution is $N_{\text{active}}/\lambda$, allowing us to interpret $1/\lambda$ as the *average revenue per active user*. We model $\log(\lambda)$ using a linear function of cohort age, age, and their interaction.

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

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.

As a preprocessing step, we standardize the features for the linear model component. 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 [7]).

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

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

Remark 4. For the linear model, we standardize the features as a preprocessing step. This standardization is omitted from the notation above for simplicity.

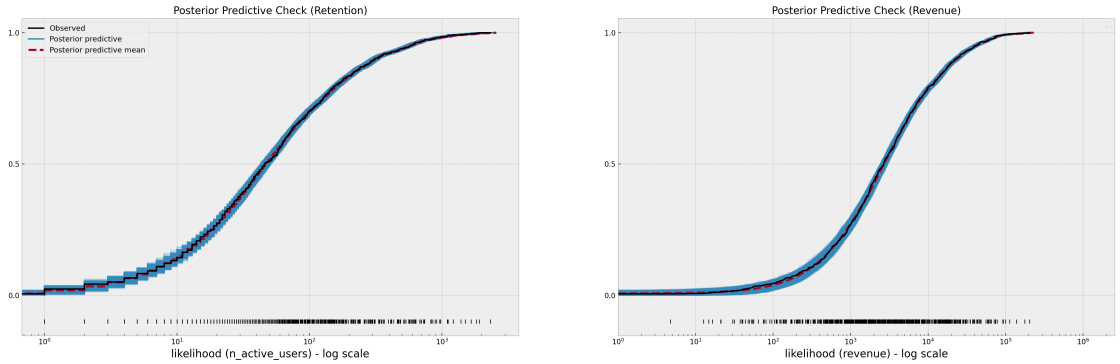

FIGURE 8. Posterior predictive distribution of the retention (left) and revenue (right) components, showing good fit to the observed data. These density plots compare the distributions of observed values (blue) with simulated values from the posterior predictive distribution (orange), providing a visual assessment of model fit. The substantial overlap between observed and predicted distributions indicates that our model successfully captures the underlying data-generating process for both retention and revenue components. This validation is crucial for ensuring that the model's uncertainty estimates are well-calibrated and that its predictions will generalize reliably to new data.

Once we have the model specification, we can implement it in PyMC (see Appendix A and [8]). Figure 8 shows the posterior predictive distribution of both model components. The trace plots for the linear terms (Figure 9) show good mixing with no divergences or convergence warnings.

4. PREDICTIONS

In this section, we present both in-sample and out-of-sample predictions from our model, demonstrating its effectiveness at capturing patterns in the data and forecasting future metrics.

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 10 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 11 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 12 displays the posterior predictive distribution compared to actual revenue values. The model successfully captures the revenue variability across different cohorts and time periods.

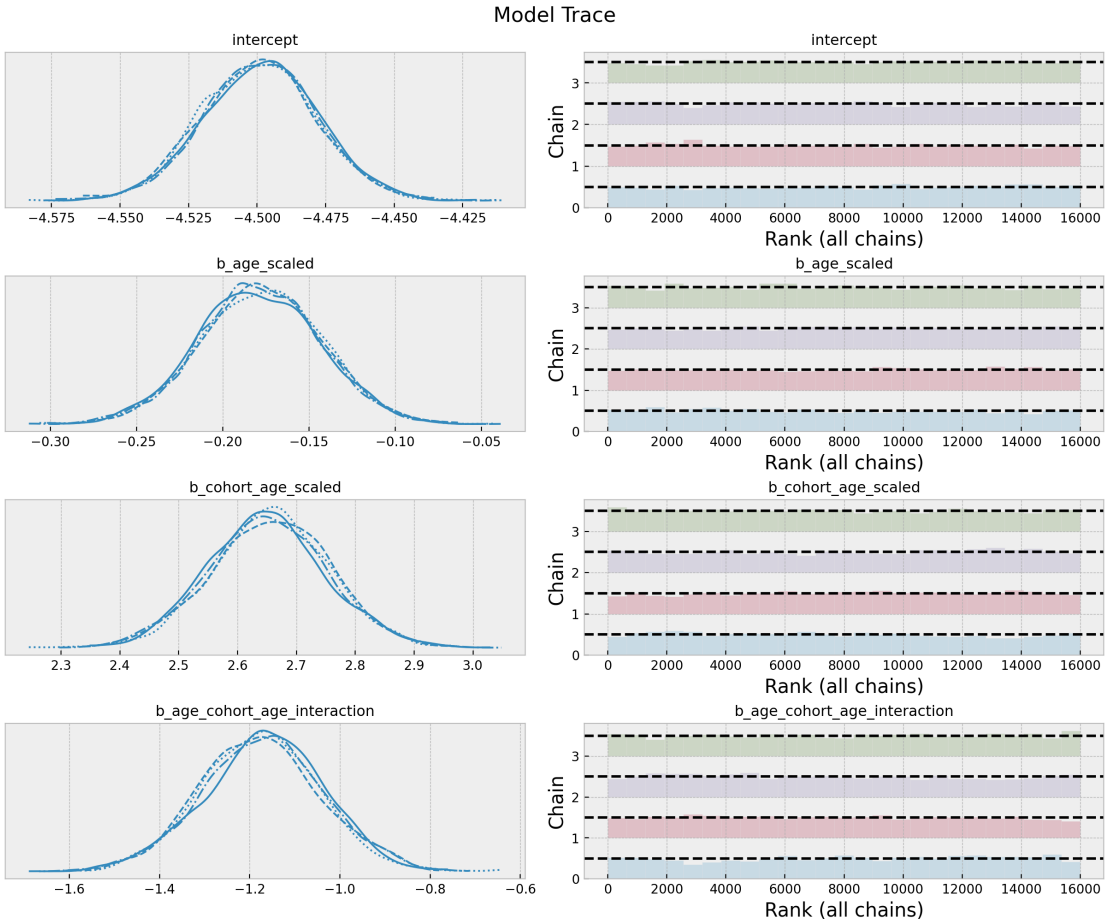

FIGURE 9. Trace plots for the linear model parameters, showing good mixing and convergence of the MCMC chains. These plots display the sampled parameter values across iterations for each coefficient in the revenue component's linear model. The rapid oscillation without visible trends or patterns indicates efficient exploration of the posterior distribution (good mixing), while the consistent range of values across multiple chains suggests they have converged to the same stationary distribution. These diagnostics are essential in Bayesian inference to ensure reliable posterior estimates, as poor mixing or lack of convergence would undermine the validity of our uncertainty quantification and predictions.

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 `2022-11-01`, which was not used during model training.

Figures 14 and 15 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:

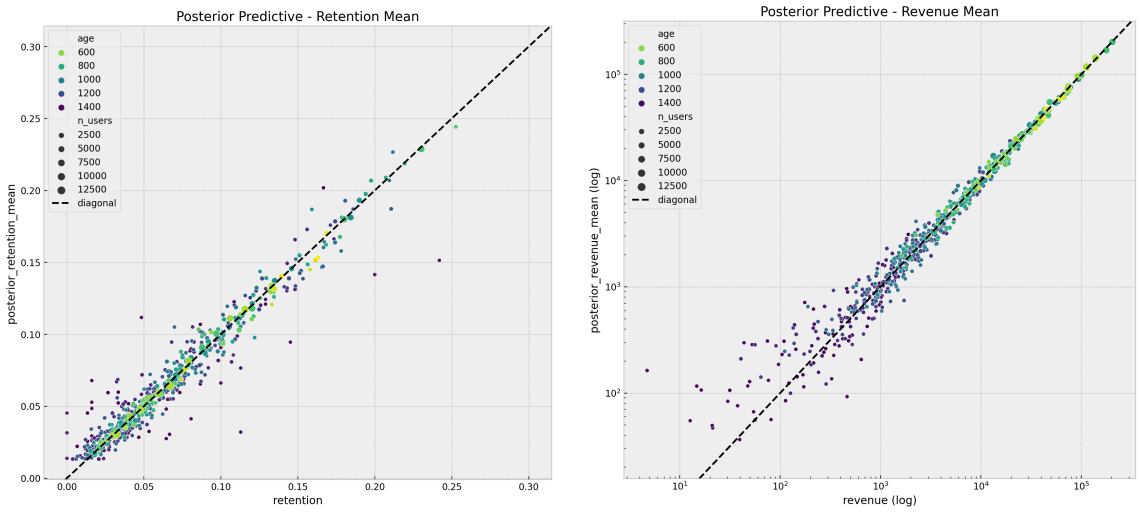

FIGURE 10. Retention (left) and revenue (right) in-sample posterior predictive mean values plotted against the actual observations. These scatter plots provide a quantitative assessment of model fit by comparing predicted versus observed values, with points closer to the diagonal line indicating better predictions. The tight clustering around the diagonal for both components demonstrates strong predictive accuracy across the entire range of values. The retention plot shows particularly high precision for larger retention rates (upper right), while the revenue plot maintains accuracy across various magnitudes, confirming the model's ability to capture both components effectively without systematic bias in specific regions.

(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 `2022-07-01` 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 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

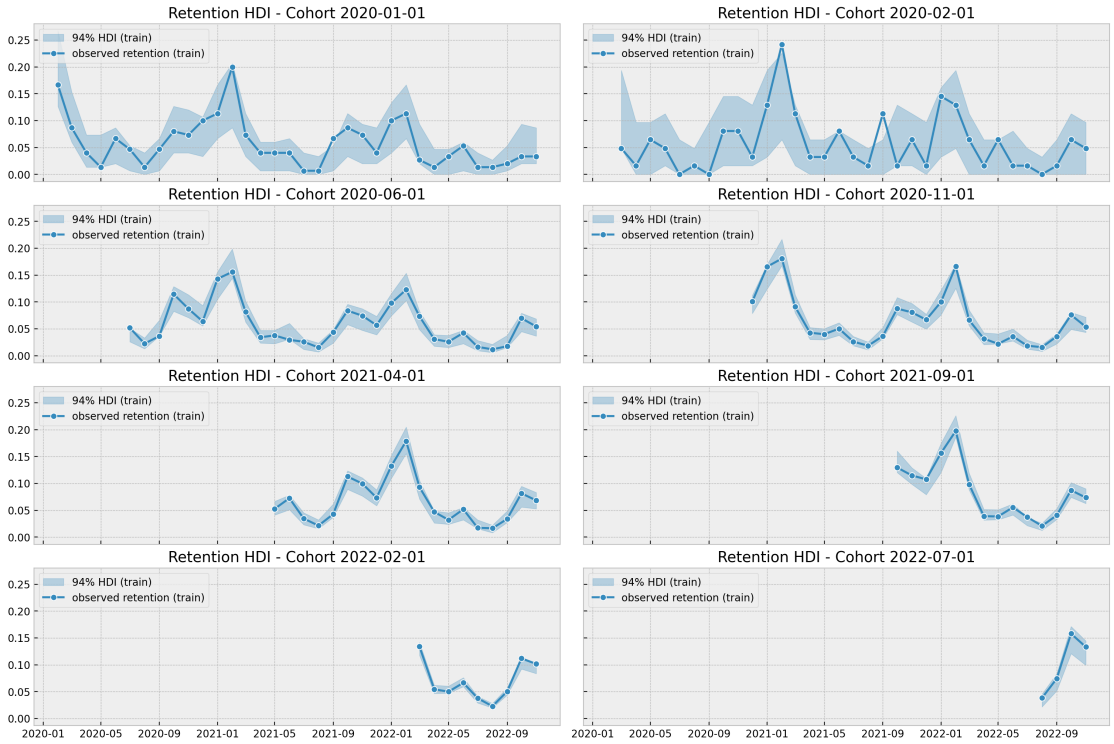

FIGURE 11. Retention in-sample posterior predictive distribution for selected cohorts, showing 94% HDI (blue shaded areas) and observed retention values (blue points). This visualization displays the model's predictive performance for retention across time for different cohorts, with uncertainty quantified 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.

decisions. The effective transfer of information across cohorts is particularly valuable for new cohorts where limited data is available.

5. Other Non-Parametric Approaches

While Bayesian Additive Regression Trees (BART) provide a powerful non-parametric approach for modeling the retention component, there are other flexible methods worth considering. In particular, neural networks coupled with efficient Bayesian inference techniques offer an alternative that combines flexibility with computational efficiency.

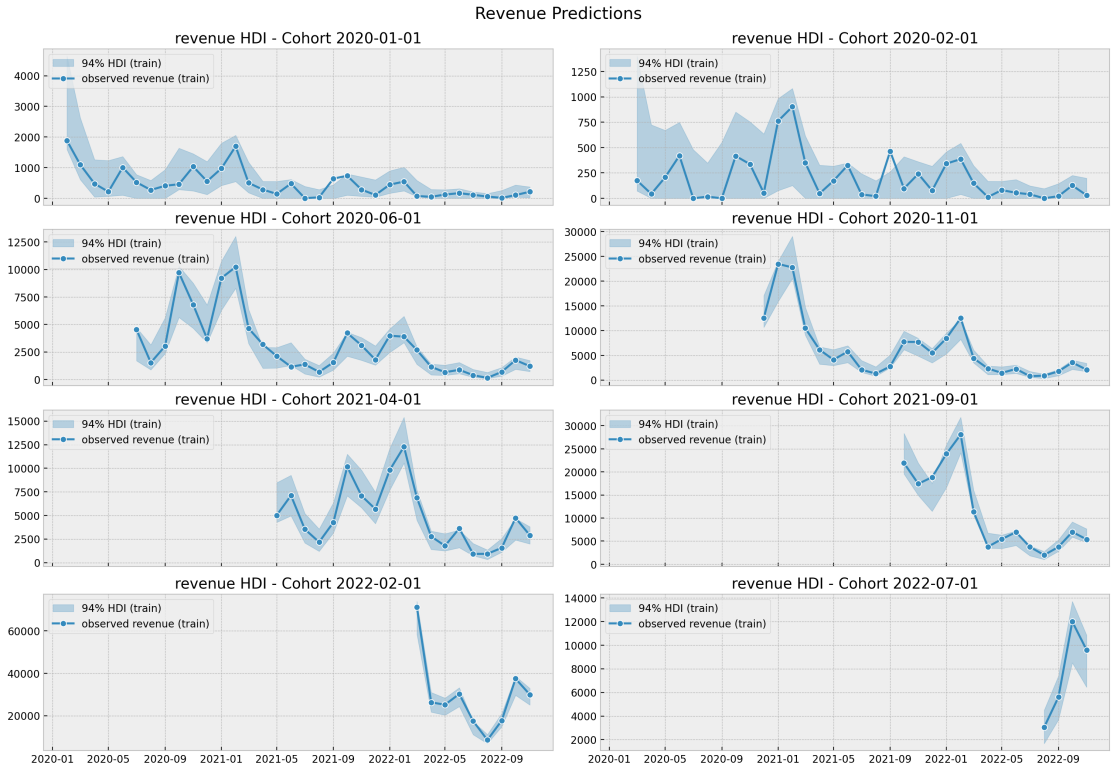

FIGURE 12. Revenue in-sample posterior predictive distribution for selected cohorts, showing 94% HDI (blue shaded areas) and observed revenue values (blue points). These plots illustrate the model's revenue predictions and associated uncertainty across time for different cohorts. The successful capture 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.

5.1. **Neural Networks with NumPyro.** As demonstrated by [10], the BART component in our model can be replaced with a neural network implemented using Flax, with inference performed using NumPyro [11]. The modified model structure becomes:

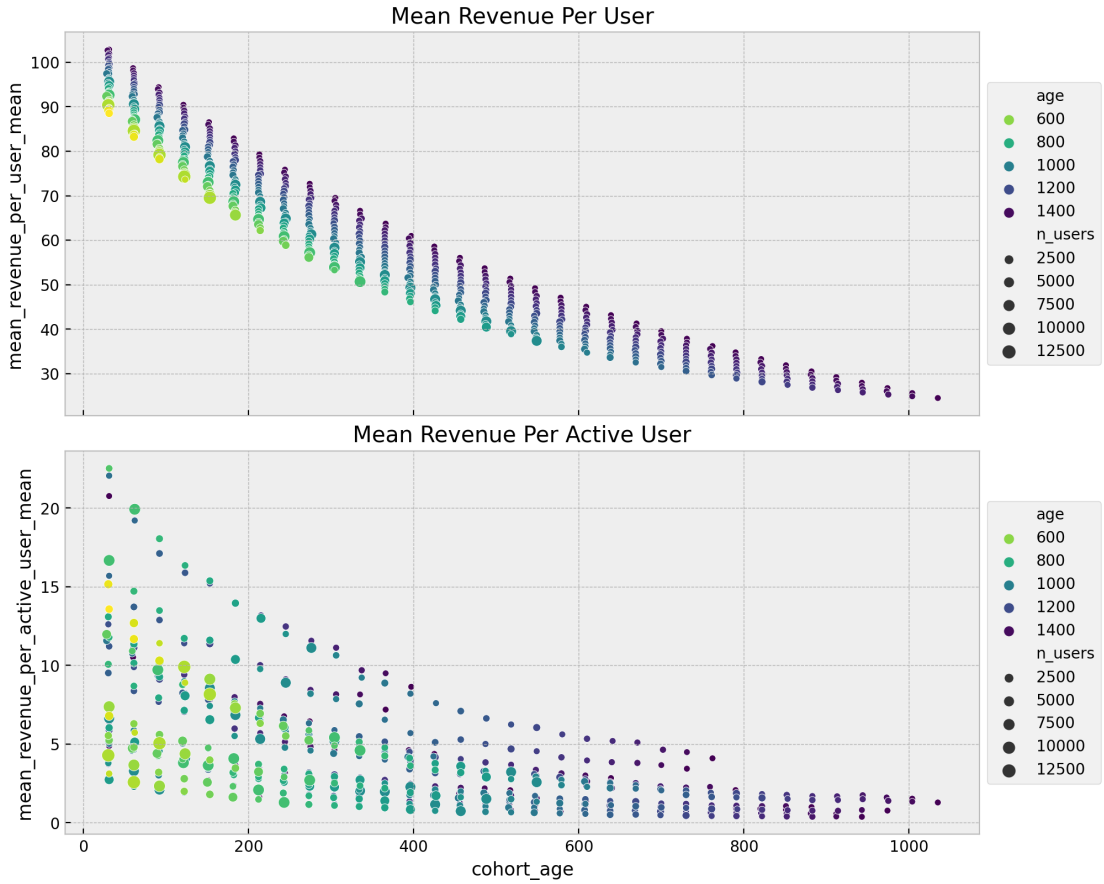

FIGURE 13. Additional view of posterior predictions across cohorts, illustrating the model's ability to capture cohort-specific patterns. This panel view organizes predictions by cohort (columns) and shows how the model adapts to the unique characteristics of each customer group. The consistent performance across cohorts of different ages and sizes demonstrates the model's robustness and the effectiveness of the partial pooling approach, where information is shared across cohorts while preserving their distinct behaviors. This balance between shared information and cohort-specific modeling is particularly valuable for businesses with diverse customer segments acquired through different channels or time periods.

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

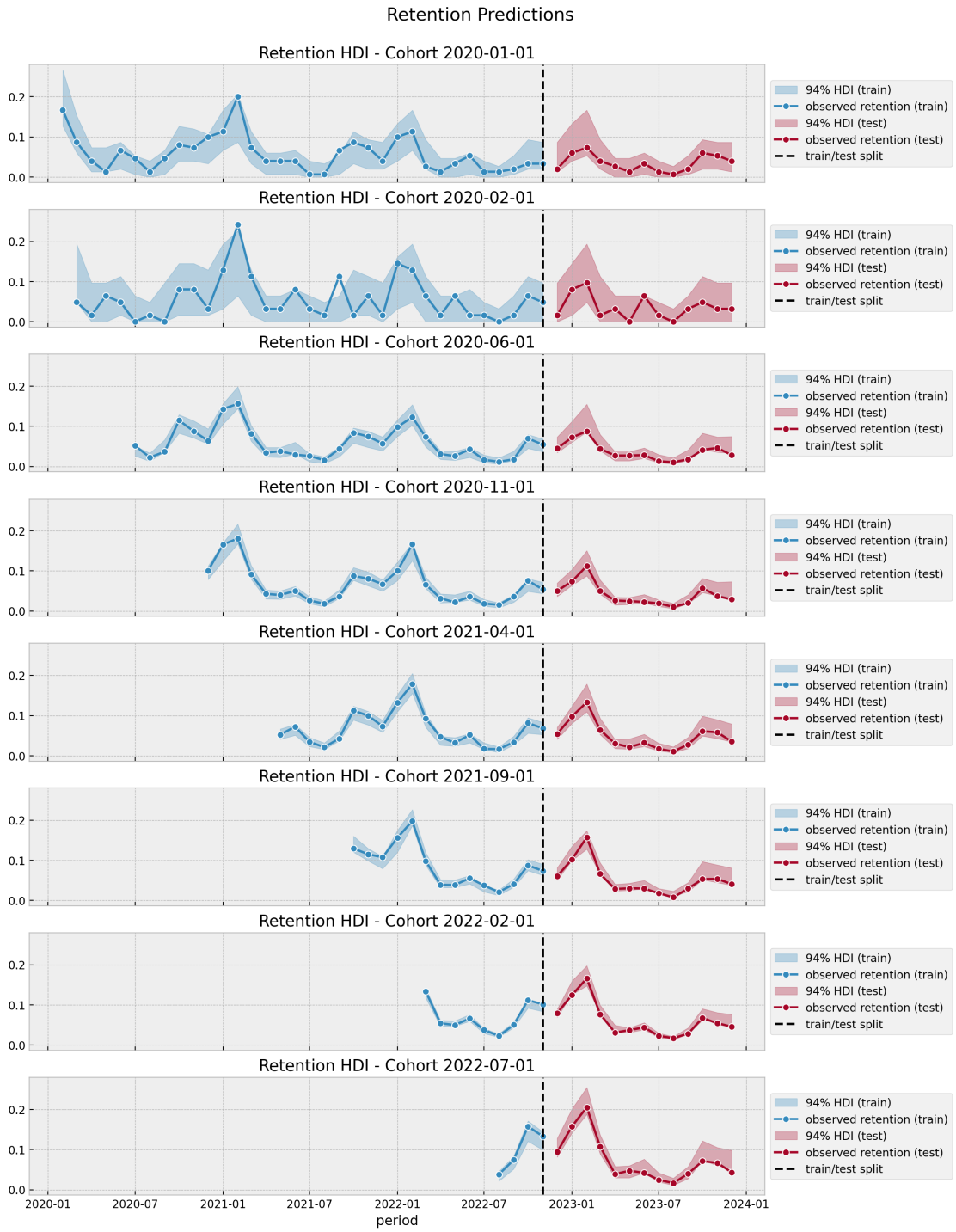

FIGURE 14. Retention out-of-sample posterior predictive distribution for selected cohorts. Blue areas represent training data 94% HDI, red areas represent test data 94% HDI, and the vertical dashed line indicates the train/test split point. This visualization demonstrates the model's forecasting capabilities by showing predictions beyond the training data period. The consistent capture of future observations within the prediction intervals validates the model's generalization ability. Notably, the prediction intervals appropriately widen as we forecast further into the future, reflecting increasing uncertainty with time horizon. The model's successful extrapolation of seasonal patterns is particularly evident, show-

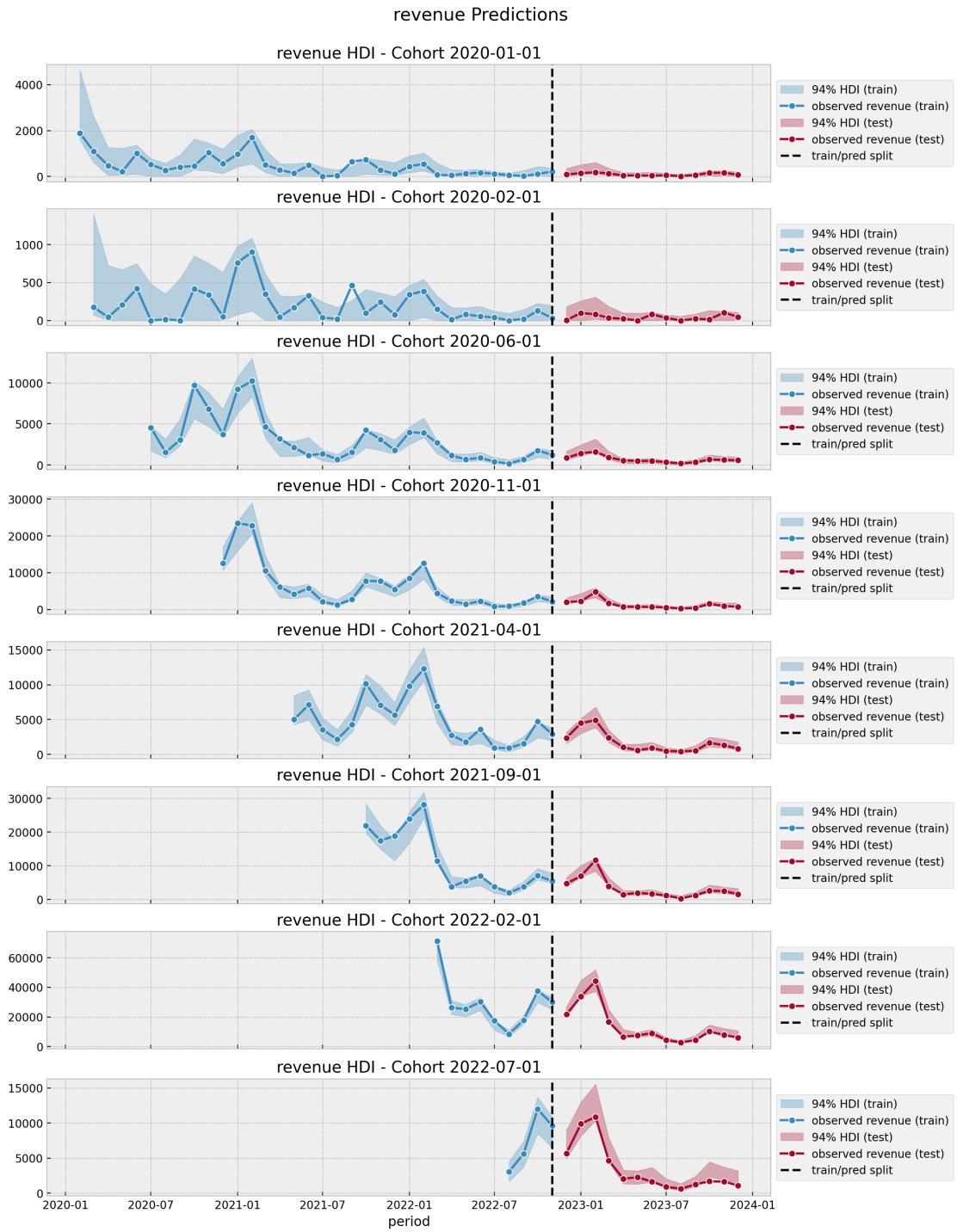

FIGURE 15. Revenue out-of-sample posterior predictive distribution for selected cohorts. Blue areas represent training data 94% HDI, red areas represent test data 94% HDI, and the vertical dashed line indicates the train/test split point. This figure evaluates the model's ability to forecast future revenue, the ultimate business metric of interest. The accurate prediction of revenue patterns in the test period demonstrates how our coupled modeling approach effectively translates retention forecasts into financial projections. The widening prediction intervals for distant forecasts provide business stakeholders with realistic assessments of financial uncertainty, enabling risk-aware decision making. For newer cohorts with

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 ([2]).
(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 [10].

6. Conclusion and Future Directions

The ability to accurately forecast retention and revenue metrics represents a significant competitive advantage in today's business environment. In this paper, we have presented a novel Bayesian approach to modeling cohort-level retention and revenue that addresses many of the limitations inherent in traditional methodologies. By combining the flexibility of Bayesian additive regression trees with the interpretability of linear models, our approach offers both analytical power and practical utility.

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.

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 several limitations that present opportunities for future research. First, while our top-down approach efficiently models cohort-level patterns, it cannot provide individual-level predictions or personalized insights. Businesses requiring customer-specific forecasts would need to complement this approach with individual-level models.

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, 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, our model currently treats cohorts as distinct entities defined solely by their start date. An interesting extension would be incorporating cohort formation factors—such as acquisition channel, initial product selection, or demographic characteristics—directly into the model structure, potentially uncovering more nuanced retention and revenue patterns.

Looking ahead, several promising research directions emerge:

(1) **Dynamic feature importance**: Developing methods to quantify how the importance of different factors affecting retention and revenue evolves over the customer lifecycle could provide valuable strategic insights.
(2) **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.
(3) **Multi-product ecosystems**: Extending the framework to handle customers who engage with multiple products or services, capturing cross-product effects on retention and spending.
(4) **Hierarchical structures**: Implementing full hierarchical Bayesian models to more formally represent the relationships between cohorts and potentially incorporate prior business knowledge.

The methodology presented in this paper represents a significant step forward in cohort-based retention and revenue modeling. By embracing the complexity inherent in customer behavior while maintaining analytical tractability, our approach bridges the gap between sophisticated statistical techniques and practical business applications. As companies continue to recognize the strategic importance of customer retention and lifetime value, flexible and robust modeling approaches like the one presented here will become increasingly essential tools in the modern business analytics toolkit.

Appendix A. Python Code

In this appendix, we present the Python code core used to implement the model in PyMC. The detailed implementation can be found in [8].

Listing 1. PyMC model implementation.

```
import pymc_bart as pmb
import pymc as pm


with pm.Model(coords={"feature": features}) as model:

    # --- Data ---
    model.add_coord(name="obs", values=train_obs_idx, mutable=True)
    age_scaled = pm.MutableData(
        name="age_scaled", value=train_age_scaled, dims="obs"
    )
    cohort_age_scaled = pm.MutableData(
        name="cohort_age_scaled", value=train_cohort_age_scaled, dims="obs"
    )
    x = pm.MutableData(name="x", value=x_train, dims=("obs", "feature"))
    n_users = pm.MutableData(name="n_users", value=train_n_users, dims="obs")
    n_active_users = pm.MutableData(
        name="n_active_users", value=train_n_active_users, dims="obs"
    )
    revenue = pm.MutableData(name="revenue", value=train_revenue, dims="obs")

    # --- Priors ---
    intercept = pm.Normal(name="intercept", mu=0, sigma=1)
    b_age_scaled = pm.Normal(name="b_age_scaled", mu=0, sigma=1)
    b_cohort_age_scaled = pm.Normal(name="b_cohort_age_scaled", mu=0, sigma=1)
    b_age_cohort_age_interaction = pm.Normal(
        name="b_age_cohort_age_interaction", mu=0, sigma=1
    )

    # --- Parametrization ---
    # The BART component models the image of the retention rate under the
    # logit transform so that the range is not constrained to [0, 1].
    mu = pmb.BART(name="mu", X=x, Y=train_retention_logit, m=50, dims="obs")
    # We use the inverse logit transform to get the retention rate
    # back into [0, 1].
    p = pm.Deterministic(name="p", var=pm.math.invlogit(mu), dims="obs")
    # We add a small epsilon to avoid numerical issues.
    p = pt.switch(pt.eq(p, 0), eps, p)
    p = pt.switch(pt.eq(p, 1), 1 - eps, p)

    # For the revenue component we use a Gamma distribution where we
    # combine the number of estimated active users with the average
    # revenue per user.
```

```
    lam_log = pm.Deterministic(
        name="lam_log",
        var=intercept
        + b_age_scaled * age_scaled
        + b_cohort_age_scaled * cohort_age_scaled
        + b_age_cohort_age_interaction * age_scaled * cohort_age_scaled,
        dims="obs",
    )

    lam = pm.Deterministic(name="lam", var=pm.math.exp(lam_log), dims="obs")

    # --- Likelihood ---
    n_active_users_estimated = pm.Binomial(
        name="n_active_users_estimated",
        n=n_users,
        p=p,
        observed=n_active_users,
        dims="obs",
    )

    x = pm.Gamma(
        name="revenue_estimated",
        alpha=n_active_users_estimated + eps,
        beta=lam,
        observed=revenue,
        dims="obs",
    )

    # --- Derived Quantities ---
    mean_revenue_per_user = pm.Deterministic(
        name="mean_revenue_per_user", var=(1 / lam), dims="obs"
    )
    pm.Deterministic(
        name="mean_revenue_per_active_user",
        var=p * mean_revenue_per_user,
        dims="obs"
    )
```

References

[1] 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. Pymc: A modern and comprehensive probabilistic programming framework in python. *PeerJ Computer Science 9* (2023), e1516.

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

[3] Fader, P., Hardie, B., and Lee, K. "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science 24* (05 2005), 275–284.

[4] Fader, P. S., and Hardie, B. G. How to project customer retention. *Journal of Interactive Marketing 21*, 1 (2007), 76–90.
[5] Fader, P. S., and Hardie, B. G. Incorporating Time-Invariant Covariates into the Pareto/NBD and BG/NBD Models. http://brucehardie.com/notes/019/, 2007.
[6] Fader, P. S., and Hardie, B. G. Fitting the sBG Model to Multi-Cohort Data. http://brucehardie.com/notes/017/, 2017.
[7] Orduz, J. Cohort Retention Analysis with BART. https://juanitorduz.github.io/retention_bart/, 01 2023.
[8] Orduz, J. Cohort Revenue & Retention Analysis: A Bayesian Approach. https://juanitorduz.github.io/revenue_retention/, 01 2023.
[9] Orduz, J. Cohort Revenue & Retention Analysis: A Bayesian Approach - Code to generate data. https://github.com/juanitorduz/website_projects/blob/master/Python/retention_data.py, 01 2023.
[10] Orduz, J. Cohort Revenue Retention Analysis with Flax and NumPyro. https://juanitorduz.github.io/revenue_retention_numpyro/, 01 2024.
[11] Phan, D., Pradhan, N., and Jankowiak, M. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv preprint arXiv:1912.11554* (2019).
[12] PyMC-Labs Development-Team. PyMC-Marketing: Bayesian marketing toolbox in PyMC. https://github.com/pymc-labs/pymc-marketing, 2023. Media Mix (MMM), customer lifetime value (CLV), buy-till-you-die (BTYD) models and more.
[13] Quiroga, M., Garay, P. G., Alonso, J. M., Loyola, J. M., and Martin, O. A. Bayesian additive regression trees for probabilistic programming, 2022.
[14] Stucchio, C. Bayesian a/b testing at vwo. https://vwo.com/downloads/VWO_SmartStats_technical_whitepaper.pdf, 2015.

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

Berlin, Germany