Quantitative Marketing and Economics
https://doi.org/10.1007/s11129-023-09272-x

A high-performance turnkey system for customer lifetime value prediction in retail brands

Yan Yan[1] · Nicholas Resnick[1]

Received: 11 August 2023 / Accepted: 3 October 2023

Abstract

Customer lifetime value (CLV) modeling underpins modern marketing analytics, enabling the development of tailored customer relationship management strategies based on the predicted future value of their customers. As part of Amperity's enterprise customer data platform (CDP), we deploy and maintain a CLV prediction system that caters to a rapidly growing list of brands across various industries, purchase behaviors, and scales. Given the impracticality of developing bespoke models for each brand, our solution must be adaptive, generalizable, and high-performing ”out of the box”. Furthermore, our platform demands daily prediction updates to facilitate prompt marketing decisions. This paper introduces a turnkey CLV prediction system that achieves state-of-the-art performance across a diverse set of brands. This system has several contributions: 1) the use of encodings and embeddings to incorporate signals from high-cardinality data; 2) a multi-stage churn-CLV modeling framework that augments additional flexibility in adjusting churn probabilities, subsequently reducing CLV prediction errors while maintaining a synergistic learning process; 3) a feature-weighted ensemble of both generative and discriminative models to accommodate diverse underlying purchase patterns. Empirical results show that our enhanced model consistently surpasses benchmark performances for twelve retail brands across six evaluation intervals from June 2020 to September 2022.

Keywords Customer lifetime value · Churn prediction · Ensemble modeling · Marketing analytics · Customer data platform

JEL Classification M31

✉ Yan Yan
yanyan@amperity.com

Nicholas Resnick
nick.resnick@amperity.com

[1] Amperity, Inc., 701 5th Ave., Seattle, WA 98104, USA

Published online: 08 November 2023

1 Introduction

Customer lifetime value (CLV) measures the revenue a business expects from a customer over a specified time period. It's a crucial metric in customer-centric marketing, as it enables businesses to enhance the long-term health of their customer relationships. Customer churn models, often a key component of CLV systems, predict which customers are likely to discontinue transacting with the business. Identifying churns is critical for many businesses since acquiring new customers often incurs greater costs than retaining existing ones. Therefore, businesses employ both CLV and churn predictions to refine marketing strategies for customer acquisition and retention, and to determine the ideal target audience for such efforts.

Amperity[1] is a Seattle-based software startup founded in 2017. Its mission is to create an intelligent customer data platform (CDP) to help businesses unify their customer data and drive better actions and insights. The foundation of Amperity's CDP is a unique entity matching system (Yan et al., 2020) that clusters customer records from disparate data sources of brands into entities representing individual consumers. Over ninety companies, spanning numerous brands, utilize Amperity to process more than fifteen billion customer records daily. These records, consolidated into almost five billion unique entities, are then merged with extensive sets of behavioral, contextual, and transactional data to form unified customer profiles. Since September 2020, Amperity has provided brands with a predictive analytics suite that offers CLV predictions based on resolved customer identities.

Constructing a CLV model atop an entity matching system presents distinct advantages. First, the incorrect identification of customer records frequently leads to errors in attributing historical expenditures across consumers. One study (Kalm et al., 2019) indicated that such misidentification predominantly affects high-value customers engaged across multiple channels and accounts. Rectifying these discrepancies empowers businesses with a more precise comprehension of their customers. Moreover, unified customer profiles connect more information about a customer, such as demographic, loyalty, browse, email engagement, and product purchase data. Although many early probabilistic models that popularized CLV modeling (Fader et al., 2005; Glady et al., 2009; Schmittlein et al., 1987) often overlooked these signals, recent studies (Chamberlain et al., 2017; Vanderveld et al., 2016; Wang et al., 2019) integrates high-dimensional features derived from behavioral customer data, frequently surpassing traditional benchmarks.

However, ensuring consistent and reliable CLV predictions for a variety of brands is considerably more challenging than performing a one-off prediction for one brand. Since its inception, our system has been employed by over seventy retail brands, with additional brands added each month. Developing bespoke models for every brand remains untenable; thus, our approach is designed to be ”turnkey”, facilitating scalability across multiple brands and over extended durations. These business requirements act as a forcing function to harden and generalize our system.

[1] https://amperity.com

1.1 High-cardinality: goldmine or haystack

While transactional data is most relevant to predicting churn and CLV, it is far from a complete picture of a customer and their relationship with a business. For instance, if a customer lives in a ZIP code where a retail brand has a large volume of stores, that will likely impact the customer's purchase propensities (Zhang et al., 2010). However, such demographic fields usually contain high-cardinality categorical data characterized by a large number of distinct values. High-cardinality features represent a major challenge for many machine learning algorithms because they provide good predictive relevance but are difficult to transform into numerical form. Binary or "one-hot" encoding performs well for low-cardinality data, but it is not practical for high-cardinality features because it produces vast and sparse datasets that require significant system memory. Clustering is utilized to reduce a large set of features to a smaller set grouped by similarity in target statistics (Johnson, 1967), and principal component analysis (PCA) (Gnanadesikan, 2011) similarly reduces categorical data to a numerical representation. However, clustering is computationally expensive on large datasets, and PCA is difficult to interpret. Johannemann et al. (2019) investigates the encoding of high-cardinality categorical variables into a condensed dimensional space using the "sufficient latent state" assumption. This approach has shown to outperform conventional techniques such as one-hot encoding, especially as the number of categories increases. However, it necessitates an added modeling step for each categorical variable. While Chamberlain et al. (2017) touches on feature embeddings in CLV modeling, the specific utilization of high-cardinality attributes have not received dedicated attention in CLV research.

We implement a feature engineering pipeline that constructs a correlation look-up table based on the empirical Bayes method introduced in Herbert (1956); Micci-Barreca (2001). We use this look-up table to map individual values in a categorical field (e.g., ZIP code and email domain) to historical purchase propensity statistics (e.g. lifetime spend and average order value). To handle more complex purchasing data with brand-specific schemas, we implement a similar Word2Vec approach described in Chamberlain et al. (2017) to convert product-level purchase data to dense embeddings. This procedure proves versatile for various categorical inputs and target outcomes, while ensuring computational efficiency and a clear interpretation of the generated features.

1.2 Churn and CLV: coupled or decoupled

Retail brands often consider marketing strategy in terms of future payback windows. Thus, marketers typically predict CLV over a specific forecasting horizon. The most commonly used horizon in retail is one year, which we adopt as the default in our system. A customer is defined as churned if they do not transact within the next horizon.

Many systems treat churn propensity and CLV prediction as two intimately coupled modeling problems. Both Vanderveld et al. (2016) and Chamberlain et al. (2017) used a multi-stage model that first predicts churn with a binary classifier and then employs a regression model to predict the CLV of non-churned customers. They further divide the CLV model into separate average order value (AOV) and order frequency (Freq) regression models. The strength of this approach is grounded in the business theory that CLV comprises customer retention and purchase subprocesses. Predicting these individual components offers actionable business context for brands. For example, businesses can identify customers with a low probability of return but a high predicted spend if they do return, initiating a high-value customer retention strategy. Venkatesan and Kumar (2004) utilized this theory in an alternative two-stage regression model that predicted order frequency and average order value. Besides additional training and tuning complexity, the main limitation of multi-stage approaches is that optimizing individual submodels doesn't necessarily optimize the final combined model. Wang et al. (2019) presented a unified modeling approach with a deep neural network (DNN) and a zero-inflated lognormal (ZILN) distribution for CLV, interpreting low or zero CLV as indicative of customers likely to churn. In our system, brands use both churn and CLV predictions, as well as average order value and order frequency, to guide various business decisions. Hence, we adopt a multi-stage modeling approach where the predicted probability of return, the predicted order frequency upon return, and the predicted average order value upon return are delivered as distinct submodel outputs, with the final CLV being the product of the three.

Churn classification for non-contractual retailers frequently faces a class imbalance issue, where few customers return for a purchase in the subsequent year. Cost-sensitive learning (CSL) and resampling techniques are common solutions to this problem (Thai-Nghe et al., 2010). CSL was applied to churn classification in Bahnsen et al. (2015) by minimizing financial costs resulting from misclassifying customers. However, in practice, quantifying such financial costs proves challenging and differs across brands. Furthermore, when training multi-stage models, a cost-minimized churn model doesn't necessarily produce the best-performing CLV model, given their distinct optimization objectives. To tackle these challenges, we developed an enhanced multi-stage CLV model. For the churn classifier, we first downsample the majority class (those who churned) during training. Subsequently, we calibrate the predicted probabilities on the original dataset to rectify the bias in posterior probability due to downsampling, providing well-calibrated predicted probabilities for churn predictions.

For the CLV model, we fit a sigmoid function to adjust the predicted probability of churn, optimizing it to minimize CLV error. This approach provides us with additional flexibility to improve CLV prediction without compromising the quality of churn predictions. It is also adaptive to the varying purchase dynamics found across retail brands. For example, the sigmoid function often adopts a different shape (especially in terms of its inflection point) for luxury brands with low purchase rates compared to budget-friendly fast fashion brands with high purchase rates.

1.3 Models: generative or discriminative

Models used for CLV modeling can generally be divided into two categories: generative probabilistic models and discriminative ML-based models. Generative models describe data-generating processes that account for observed data using a few model parameters. The initial evolution of generative CLV modeling comprised "Buy Till You Die" (BTYD) parametric models, such as the Pareto/NBD (Schmittlein et al., 1987), which presumes a Pareto-II distribution for customer lifetime and a negative-binomial distribution (NBD) for purchase frequency. The Pareto distribution is substituted with a Beta-geometric distribution in Fader et al. (2005), assuming each customer has a fixed probability to churn after a purchase. Markov Chains have also been recognized as a valid CLV methodology since they are model-free and don't necessitate assumptions about purchase or churn behavior. An thorough review of ten different probabilistic models is given in Jašek et al. (2019). They concluded that nearly all Pareto/NBD models consistently deliver strong results across various brands and evaluation metrics (e.g., MAE, Spearman's correlation coefficient, sensitivity across quantiles). However, Markov Chains not only performed worse than P/NBD but also fell behind a simple status-quo baseline, which assumes that a customer's purchase behavior in the next period will be the same as in the previous one. Typically, the strength of probabilistic models is twofold: they offer relative clarity in business and marketing contexts, especially when applying separate distributions to churn and purchase frequency, and they are trained on complete historical data, eliminating the need to hold out a period of time to generate target variables.

Discriminative ML models, such as Random Forest (RF) and Deep Neural Networks (DNNs) offer more flexible and powerful approaches to fitting many parameters to data. As the amount of feature data and computational resources increased, they emerged as robust solutions for predicting CLV. Evidence was first presented in Gupta et al. (2006) that RF models outperformed previous state-of-the-art BTYD models. The RF approach was improved by including customer engagement features from app and browse data (Vanderveld et al., 2016), and Martínez et al. (2018) extended the tree-based models with extreme gradient boosting (Chen & Guestrin, 2016) in a set of non-contractual purchasing cases in retail. Chamberlain et al. (2017) also showed a DNN with enough hidden layers achieved comparable performance to RF. Seq2Seq networks leveraged the time-series nature of customer transaction data and showed competitive results to existing methods on retail datasets, despite more expensive training costs (Bauer & Jannach, 2021).

Contrary to generative models, most discriminative ML methods aren't reliant on distributional assumptions. However, these models come with greater computational demands, offer limited interpretability, and have strict requirements for training data generation, especially in time-series contexts. A duration equal to the prediction horizon must be reserved for generating target variables when creating training data for a discriminative model (see Fig. 1 for comparison). For a prediction horizon of one year, the data from the previous year is held out for target variable generation, thus constraining the amount of historical data available for model training.

Blending multiple models tends to improve CLV prediction performance. Bauer and Jannach (2021) extended their work on Seq2Seq models by stacking them with a gradient-boosted machines model. Validation on the same online retail datasets indicated the models captured different underlying patterns, improving overall performance when ensembled. While not in retail, Desirena et al. (2019) also found stacking DNNs performed well for maximizing CLV in insurance businesses.

To the best of our knowledge, the majority of modeling approaches for churn and CLV remain a single model, and ensembles are exclusively composed of multiple discriminative models. Our approach, on the other hand, utilizes a novel ensemble

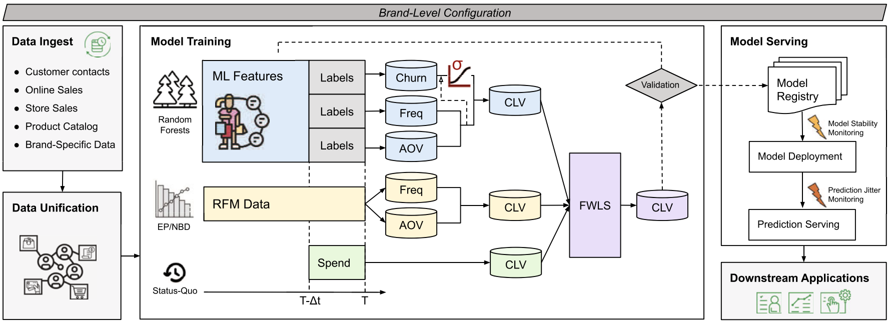


Fig. 1 An overview of our CLV prediction system

approach that blends predictions from both generative and discriminative models. We acknowledge the importance of both: Generative probabilistic models impose strict assumptions that, while sometimes violated, have proven to be reasonably accurate in a non-contractual retail setting. However, when purchasing patterns are highly irregular and rich behavioral datasets are available, modern ML techniques often produce stronger predictive performance. Our ensemble model includes a Random Forest model with rich features, an extended Pareto/NBD (EP/NBD) model, and a status-quo baseline model (Jašek et al., 2019). The three heterogeneous base models are ensembled using feature-weighted linear stacking (Sill et al., 2009). This approach combines the flexibility and richness of ML techniques with the long history and statistical grounding of probabilistic approaches.

The key contributions of this work include:

1. We apply empirical Bayesian-based encoders and use Word2Vec embeddings to generate rich features beyond transaction data, especially for categorical variables with high cardinality and high sparsity.
2. We develop an enhanced multi-stage churn-CLV model that applies a learned sigmoid function to adjust the churn probability, thereby minimizing CLV errors resulting from churn misclassifications.
3. We implement a feature-weighted linear stacking model that leverages both generative and discriminative models. This ensemble model consistently outperforms its base models across varying evaluation scenarios.

The rest of the paper is organized as follows. Section 2 provides a system design overview. Section 3 discusses our featurization techniques. Section 4 introduces our enhanced multi-stage churn-CLV model. Section 5 reviews feature-weighted linear stacking and explains our implementation. Section 6 presents empirical evidence of our approach's effectiveness. Section 7 shares our learning and concludes our discussion.

2 System overview

We design our CLV system to serve as a turnkey solution for many brands across various industries. Our requirements for a high-performance turnkey solution include

1) the flexibility to incorporate new input data fields, 2) minimal manual feature engineering and model tuning, 3) adaptiveness to brand-specific purchase patterns, 4) reliable predictions, and 5) operational stability. Section 1 introduces the challenges we encountered when designing and implementing this system. This section provides an overview of our solution. The system architecture is shown in Fig. 1.

Amperity's data architecture runs atop infrastructure as a service (IAAS), such as Amazon AWS or Microsoft Azure. The entire data processing pipeline is implemented on Apache Spark for high throughput and scalability. Brands send us daily data updates, which trigger the data unification pipeline. Records representing the same customer are merged together to create a unified profile with all customer attributes and interactions with a brand.

Our machine learning module is packaged in an internal Python library named AmpLearn, installed on Spark clusters and built, tested, and deployed via CI/CD pipelines. While the module itself is brand-agnostic, it is applied to each brand via a robust configuration layer. A brand's configuration contains five types of information: input data schema, preprocessing requirements, model settings, output data schema, and production job settings. During model training, a brand's datasets are first read from cloud storage and transformed into a standardized input format according to their input data schema. Then a training set is formed by combining handcrafted transactional features with encodings and embeddings for high-cardinality categorical data. The training set is fed into an ensemble CLV model which comprises of: 1) a multi-stage Random Forest model (M-RF) subdivided into churn, order frequency, and average order value submodels, 2) an EP/NBD model, and 3) a status-quo (or business-rule) model. These three models are integrated via a feature-weighted linear stacking ensemble to produce the final CLV predictions, which are then delivered to brands in their preferred output data schema.

Models are retrained weekly to incorporate new signals with reasonable computational cost. Model artifacts are registered with MLflow for easy storage, registration, and rollback. Once a brand is activated in our CLV system, we monitor both weekly retrained models and daily predictions to ensure the reliability of predictions delivered to brands. Two types of data drift are measured:

1. Model stability (monitored weekly): $\Delta(\mathrm{Pred}(D_i, M_j), \mathrm{Pred}(D_i, M_{j+1}))$. The difference in predictions by applying different model versions j and $j+1$ to the same data version i.
2. Prediction jitter (monitored daily): $\Delta(\mathrm{Pred}(D_i, M_j), \mathrm{Pred}(D_{i+1}, M_j))$. The difference in predictions by applying the same model version j on different data versions i and $i+1$.

Here, we use i and j as notations to represent respective versions of the data and the model. The predictions corresponding to data version D_i and model version M_j are denoted by $\mathrm{Pred}(D_i, M_j)$. We use the metric $\Delta(\cdot)$ to measure the drift between predictions of different versions. Commonly adopted metrics for this purpose include the Kullback-Leibler Divergence and the difference in means. If a substantial drift is observed, it prompts alerts necessitating operator examination and potential intervention. In the absence of such drift, the model proceeds to deployment, and predictions are served.

Most current CLV research is centered on single-brand, one-time delivery scenarios. In contrast, our system: 1) extends our ML pipeline to multiple brands across diverse industries, and 2) provides daily predictions as as part of the service level agreement (SLA). Our CLV predictions power marketing campaigns; therefore, unreliable or volatile predictions can have significant downstream impacts for our customers.

3 Features

After the data unification process, canonicalized customer demographic information, order-level transactions, and product details are available across all brands with consistent schemas. Supplementary data fields, such as in-store information, email engagement, event participation data, and others, vary in availability between brands. On average, 65% of the data fields contain high-cardinality categorical attributes, and 20% are brand-specific fields. When combined with data variability across brands, a scalable and generalizable feature engineering process becomes increasingly important. In this section, we present techniques designed to address these challenges.

3.1 Empirical bayesian encoder

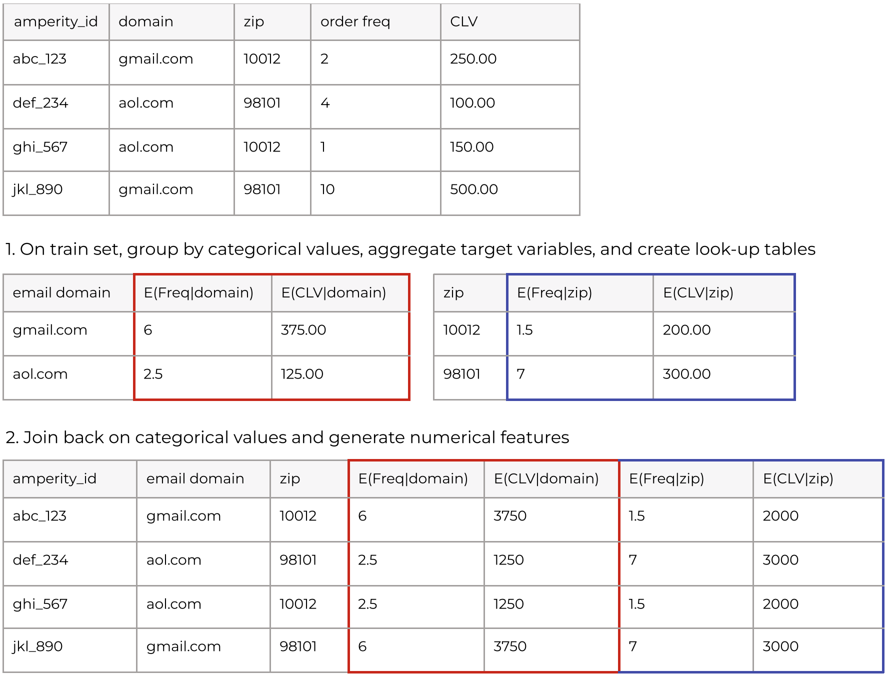
In addressing high-cardinality features, we employ a simple yet elegant technique that transforms each categorical value into a numerical representation based on its correlation with the target variable. This technique is anchored in the empirical Bayes (EB) statistical method, as proposed by Herbert (1956) and generalized as a data preprocessing paradigm by Micci-Barreca (2001). However, its application in CLV research remains less explored. The encoder computes the conditional expectation of a target variable θ given a particular value X_i of a high-cardinality categorical variable X, expressed as:

$$f_{\mathrm{EB}}(X_i) = E(\theta | X = X_i) = \frac{\sum_{k \in L_i} \theta_k}{n_i}, \tag{1}$$

where L_i denotes the set of observations corresponding to the value X_i and n_i represents the sample size. For binary targets, the formulation in Eq. 1 remains largely consistent, but the expected value is replaced by the estimated probabilities, meaning $\sum_{k \in L_i} \theta_k$ converts to the count of positive observations (see Fig. 2 for a toy example).

The EB encoder allows us to encode any high-cardinality categorical feature as a continuous scalar feature. It handles low frequency values and missing values very well[2]. The features are simple to interpret, inspect, and monitor. Their predictive relevance for new fields can be automatically captured without bespoke feature engineering. Moreover, they can be implemented easily using database queries. Addi-

[2] A weighting factor represented as a function of the sample size should be used to blend $E(\theta|X = X_i)$ with the sample expectation $\bar{\theta}$, i.e., $f_{\mathrm{EB}}(X_i) = \lambda(n_i)E(\theta|X = X_i) + (1 - \lambda(n_i))\bar{\theta}$. An intuitive choice for $\lambda(n_i)$ as discussed in Micci-Barreca (2001) is $n_i/(\sigma_i^2/\sigma^2 + n_i)$, where σ_i^2 is the variance given $X = X_i$ and σ^2 is the variance of the entire sample. Noisier (higher variance) data in the sample compared to the overall dataset results in smaller $\lambda(n_i)$ and more shrinkage toward the population mean.



amperity_id	domain	zip	order freq	CLV
abc_123	gmail.com	10012	2	250.00
def_234	aol.com	98101	4	100.00
ghi_567	aol.com	10012	1	150.00
jkl_890	gmail.com	98101	10	500.00

1. On train set, group by categorical values, aggregate target variables, and create look-up tables

email domain	E(Freq|domain)	E(CLV|domain)
gmail.com	6	375.00
aol.com	2.5	125.00

zip	E(Freq|zip)	E(CLV|zip)
10012	1.5	200.00
98101	7	300.00

2. Join back on categorical values and generate numerical features

amperity_id	email domain	zip	E(Freq|domain)	E(CLV|domain)	E(Freq|zip)	E(CLV|zip)
abc_123	gmail.com	10012	6	3750	1.5	2000
def_234	aol.com	98101	2.5	1250	7	3000
ghi_567	aol.com	10012	2.5	1250	1.5	2000
jkl_890	gmail.com	98101	6	3750	7	3000

Fig. 2 Example computation of empirical Bayesian encoders summarizing order count and spend for two high-cardinality categorical features: email domain and zipcode

tionally, the computation is fast and parallelizable, making it well-suited for large-scale environments.

3.2 Word2Vec embedding

While the EB encoder relates purchase propensities to high-cardinality categorical attributes, it does not necessarily capture complex purchasing patterns in the data. In recent CLV research, neural embeddings (Mikolov et al., 2013) have emerged as a popular method for generating dense numerical features from such patterns. This is especially true for large datasets, such as itemized browse data (Chamberlain et al., 2017), which usually contain rich and continuously evolving product-level information.

We adopt an approach similar to that described in Chamberlain et al. (2017), but we utilize product-level purchase data instead of browse data. Itemized transaction data is grouped at the product level, and customers who have purchased that product are sorted in ascending order by purchase time. In the context of Word2Vec's typical application in natural language processing, we treat products as documents and customers (represented by ID strings) as words. Similar to the Word2Vec assumption that words appearing in close proximity are related, we infer that customers who purchase a specific product around the same time have similarities. The result of the Word2Vec process is a customer-level embedding, which we directly use as features in our model.

When training Word2Vec, we use data up to $T - \Delta t$ (see Fig. 1, where T represents the inference time and ΔT denotes the forecasting window). A challenge in applying Word2Vec in a time-series context is the lack of a straightforward method to refresh embeddings at the inference time. Chamberlain et al. (2017) addresses this challenge, and we adopt a similar solution. During inference, we gather the customers who have bought each product and compute the average of their embeddings to derive product-level embeddings. Initially, we calculate product-level embeddings by averaging the embeddings of customers who have purchased the product. Then, for customers present as of the training time, we average their initial embedding with new product embeddings from their subsequent purchases. For new customers, we simply set their embedding as the mean of the embeddings from the products they have purchased.

3.3 Handcrafted features

In addition to features associated with RFM data, such as recency (time since last purchase), frequency (number of purchases), and monetary value (average order value) commonly used in BTYD models, and the features derived via empirical Bayesian encoder and Word2Vec embedding as previously discussed, we also include several handcrafted features that consistently exhibit strong predictive relevance across brands. These features serve as the default set, offering a robust performance baseline. These are some examples:

- **Clumpiness**: A metric that quantifies the irregularity of a customer's inter-purchase times, defined as the ratio between the days spanning the first and last purchases and the days since the first purchase.
- **Holiday purchases**: The proportion of a customer's purchases done during holidays relative to non-holidays.
- **Discount/return/cancellation tendency**: Features associated with discounted, returned, and canceled purchases.
- **Multi-channel purchases**: The spread of a customer's purchases across different channels.
- **Email engagement**: Metrics like the number of email opens and clicks, and the recency of their last email engagement.
- **Seasonality**: The month and year for which the training sample was produced (samples are created in monthly intervals starting from the most recent period and extending back to two years prior.)

A feature ablation study, depicted in Fig. 3, illustrates the significance of each feature group. Features were divided into three categories: those generated by empirical Bayesian encoders, by Word2Vec embeddings, and all other transactional features. The model was trained with every combination of the three categories, and its performance was evaluated using MAE on a test set. The inclusion of both Word2Vec embeddings and Bayesian encoders has shown a 7.42% reduction in MAE.

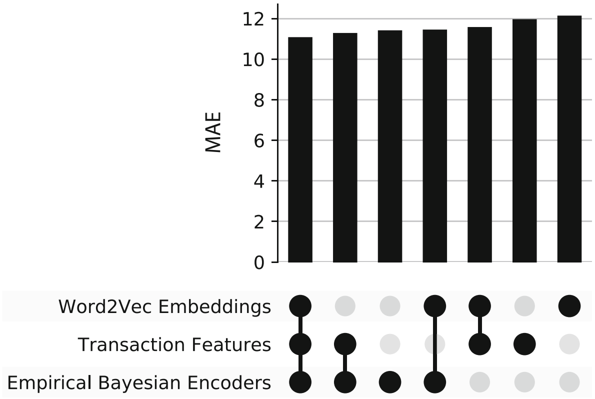


Fig. 3 An ablation study illustrating the significance of each feature group, evaluated using the mean absolute error (MAE) of the Customer Lifetime Value (CLV) prediction, measured in dollars

4 An enhanced multi-stage model

Our discriminative model is developed in multiple stages. First, a binary classifier is built to predict the probability that a customer will churn in the next time window. Subsequently, a CLV regression model is trained on customers predicted to return. This CLV model is a product of both an average order value (AOV) regression model and a purchase frequency (Freq) regression model. In our system, delivering these submodels is advantageous because they assist brands in guiding different marketing actions. The predicted CLV for a given customer x is denoted as $\mathrm{CLV}(x)$. The multi-stage model can be described as follows:

$$\begin{aligned} \mathrm{CLV}(x) &= \mathrm{Prob}_{\mathrm{return}}(x)\ \mathrm{CLV}_{\mathrm{return}}(x) \\ &= \mathrm{Prob}_{\mathrm{return}}(x)\ \mathrm{AOV}_{\mathrm{return}}(x)\ \mathrm{Freq}_{\mathrm{return}}(x), \end{aligned} \tag{2}$$

Where the probability of a customer returning and the probability of a customer churning are inversely related, such that $\mathrm{Prob}_{\mathrm{return}} = 1 - \mathrm{Prob}_{\mathrm{churn}}$. Here, we assume that the CLV of a customer who has churned, or $\mathrm{CLV}_{\mathrm{churn}}$, is zero.

4.1 Interconnected concerns of CLV and churn

The combined churn-CLV model is widely used in many CLV systems because the two problems are interconnected. For simplicity, some systems (Chamberlain et al., 2017; Vanderveld et al., 2016) treat $\mathrm{Prob}_{\mathrm{churn}}$ in Eq. 2 as a boolean input. There are several challenges with this framework that are often overlooked:

- For non-contractual businesses, the two classes, return versus churn, tend to be imbalanced (see the retention rate in Section 6.1 for examples). In the face of highly imbalanced data, many classifiers are biased by the majority class examples, leading to a high false-negative rate. Techniques such as undersampling the majority class or resampling the minority class can help address this problem, but they also alter the priors of the training set, influencing the posterior probabilities of a classifier.

- Many classifiers assume that the costs of misclassification (both false negatives and false positives) are equivalent. In practical scenarios, this assumption is seldom accurate. For instance, the cost associated with mistakenly predicting a customer will churn when they don't is typically less than the cost of failing to predict the departure of a loyal customer.
- The costs related to prediction errors in churn and CLV models differ. A churn model, even if well-adjusted for class imbalance, may not also minimize the CLV prediction error. This discrepancy arises because different types of churn misclassifications influence CLV errors to varying degrees. Through empirical observation, we found this issue is especially pronounced in brands with high AOVs and elevated churn rates.

4.2 Our algorithm

We propose an enhanced multi-stage modeling framework to address these issues (see Algorithm 1). Our algorithm aims to minimize the impact of errors resulting from incorporating a churn model into CLV calculations. In our approach, we use the in-sample error of CLV predictions to fit a sigmoid function. This provides an additional flexibility, allowing for independent adjustment of the churn probability to minimize the CLV error.

Algorithm 1 Enhanced Multi-Stage Churn-CLV Model.

```
Input: Dataset $D$
Output: $\text{Prob}_{\text{return}}, \text{AOV}_{\text{return}}, \text{Freq}_{\text{return}}, \text{CLV}$
1: $D_{\text{train}}, D_{\text{test}} \leftarrow$ split $D$ by time
2: $D_{\text{train}_{\mathbf{B}}} \leftarrow$ balance classes on $D_{\text{train}}$
3: $\text{Prob}_{\text{return},\mathbf{B}} \leftarrow$ train a Random Forest classifier on $D_{\text{train}_{\mathbf{B}}}$
4: $\text{Prob}_{\text{return}} \leftarrow$ calibrate $\text{Prob}_{\text{return},\mathbf{B}}$ on $D_{\text{train}}$
5: $\text{AOV}_{\text{return}}, \text{Freq}_{\text{return}} \leftarrow$ train Random Forests on $D_{\text{train}}$
6: $\text{CLV}_{\text{return}} \leftarrow \text{AOV}_{\text{return}}\ \text{Freq}_{\text{return}}$
7: $\sigma_{t_1^*,t_2^*} \leftarrow \arg\min_{t_1,t_2} \left|\sum \sigma_{t_1,t_2}(\text{Prob}_{\text{return}})\text{CLV}_{\text{return}} - \sum \text{CLV}_{\text{Actual}}\right|$ on $D_{\text{train}}$ with cross-validation
8: $\text{CLV} \leftarrow \sigma_{t_1^*,t_2^*}(\text{Prob}_{\text{return}})\text{CLV}_{\text{return}}$
9: Evaluate $\text{Prob}_{\text{return}}$ and CLV on $D_{\text{test}}$
```

Similar to Platt scaling (Platt et al., 1999), we use a sigmoid function to transform the probability output of the churn classifier. However, the objectives differ. While Platt's method is used for probability calibration (as seen in Step 4 of Algorithm 1), our approach aims to minimize the impact of CLV errors due to churn misclassifications. The larger the values of t_1 and $|t_2 - 0.5|$ (with 0.5 as the default classifier threshold), the more distortion the sigmoid function introduces. Figure 4 presents examples of estimated (t_1^*, t_2^*) values for three retail brands (see descriptions of the datasets in Section 6.1) and illustrates how the CLV errors vary with t_2. Among these brands,

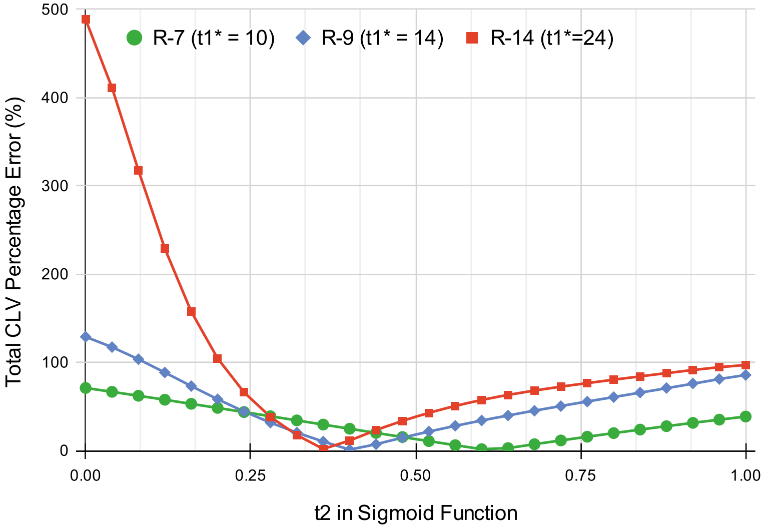


Fig. 4 Examples of (t_1^*, t_2^*) in three retail brands. The total CLV percentage error is a normalized version of the total revenue error, divided by the total number of customers

R-7 has the lowest AOV ($82.9) and the highest return rate (31.3%), while R-14 has the highest AOV ($188.5) and the lowest return rate (8.3%). R-14 undergoes the most significant adjustment, with t_2 dropping to as low as 0.28. We employ the total revenue error (i.e., $\left|\sum \sigma_{t_1,t_2}(\text{Prob}_{\text{return}})\text{CLV}_{\text{return}} - \sum \text{CLV}_{\text{Actual}}\right|$, where $\sum \text{CLV}_{\text{Actual}}$ represents the actual CLV of all customers) as the cost function. By accurately predicting total revenue, the model better captures the overall purchase pattern and is less prone to overfitting compared to individual-level metrics, such as MAE. Nevertheless, this approach consistently reduces MAE as well (as detailed in Section 6). Apart from CLV errors, other financially-based cost functions can be also employed to enhance various business objectives.

5 Ensemble learning

One of the key challenges in using discriminative models for CLV modeling is the necessity to reserve recent data to compute the target variable for training, a requirement not present in generative models. To mitigate this issue, we introduced a unique ensemble of models that combines our enhanced multi-stage Random Forest model (EM-RF) with an EP/NBD model and a status quo (SQ) model. The ensemble model shows consistent improvement over all three base models, both across different brands and over various evaluation periods.

5.1 Base models

The ensemble is comprised of the following three heterogeneous base models:

EM-RF model This enhanced multi-stage Random Forest model leverages the featurization techniques discussed in Section 3 and incorporates an additional churn probability adjustment function to reduce the CLV prediction error as described in Section 4. We use Random Forests as they are well-suited for distributed computing, resistant to overfitting, and have historically shown strong performance in CLV modeling (Chamberlain et al., 2017; Vanderveld et al., 2016).

EP/NBD model The Extended Pareto/NBD model (EP/NBD), described in Schmittlein et al. (1987) and Gupta et al. (2006), has exhibited state-of-the-art performance according to an empirical benchmark (Jašek et al., 2018). Typically, the Pareto/NBD model is paired with an additional model for predicting AOV, and we employ the Gamma-Gamma extension (Fader & Hardie, 2013) for this purpose. The derivation of the EP/NBD is omitted for brevity, and can be found in Fader and Hardie (2005) and Gupta et al. (2006). This derivation indicates that a customer's observed purchase frequency, the interval between the first and last purchases, and the total observation time are the sole features necessary to train the Pareto/NBD. Likewise, the Gamma-Gamma extension utilizes purchase frequency and average order value features. Thus, only four distinct transactional features are required to fit this model. Being a generative model, it eliminates the need to hold out a specific time period for generating target variables in the training data, which is beneficial for long-term time-series forecasting as the model can utilize the entire transaction history.

SQ model We also include a status-quo model (Jašek et al., 2019) that presumes the behavior of each customer in the upcoming period will mirror their behavior from the previous period. In other words, their projected CLV for the subsequent year equals their total spend from the preceding year. Despite its simplicity, this model captures the order value distribution and serves as a reliable baseline when no better information is available. It remains a prevalent business rule in scenarios where a sophisticated predictive model is absent.

5.2 Feature-weighted linear stacking

Since its introduction by Sill et al. (2009), feature-weighted linear stacking (FWLS) has been shown to enhance predictive performance beyond that of individual base models. Unlike standard linear stacking, which blends base models with constant weights, FWLS assumes that the predictive power of each base model varies based on individual-level information, or meta-features. For instance, EP/NBD might be more reliable than an RF model for customers with an extensive and consistent transaction history. FWLS retains many advantages of linear models, such as low computational costs, minimal tuning, and interpretability, while still offering a substantial improvement in predictive performance.

Given a customer x, the ensembled CLV prediction using FWLS is expressed as:

$$\mathrm{CLV}_{\mathrm{fwls}}(x) = \sum_{k=1}^{K} \sum_{m=1}^{M} v_{m,k} f_m(x) \times \mathrm{CLV}_k(x). \tag{3}$$

Here, $\mathrm{CLV}_k(x)$ denotes the prediction of the k^{th} base model for customer x. The weights assigned to these predictions, $v_{m,k} f_m(x)$, are linear functions of the customer's meta-features $f_m(x)$. Specifically, $v_{m,k}$ represents the coefficients or significance values corresponding to each meta-feature for each base model. Thus, the FWLS optimization problem is equivalent to fitting a linear regression model with $K \times M$ features. In its essence, FWLS acknowledges that different base models might possess varying predictive strengths for distinct customer segments. By permitting the blending weights to be influenced by meta-features, FWLS adapts to each customer's unique characteristics, potentially leading to enhanced and more precise predictions.

While adding more meta-features could potentially enhance predictive performance, it's advantageous to keep a limited set when deploying FWLS in a production environment because the computational training cost increases quadratically with the number of meta-features. Furthermore, we refrain from creating bespoke meta-features that depend on brand-specific data, as these often don't provide universally applicable performance boosts. Through cross-validation on a validation set, we identified a compact set of ten meta-features. These were chosen for their consistent performance enhancement and their reliance solely on basic transactional data, which is universally available across brands. The selected meta-features are as follows:

- f_1 = frequency over the last 3 months,
- f_2 = frequency over the last 6 months,
- f_3 = order frequency over the last 12 months,
- f_4 = lifetime order frequency,
- $f_5 = \frac{\text{order frequency over the last 12 months}}{\text{lifetime order frequency}}$,
- $f_6 = \frac{\text{days since most recent order}}{\text{days since first order}}$,
- $f_7 = \frac{\text{average order value over the last 12 months}}{\text{lifetime average order value}}$,
- f_8 = clumpiness,
- f_9 = average order discount percentage,
- f_{10} = number of holiday orders.

5.3 Implementation considerations

There are several challenges associated with applying FWLS to our CLV regression problem, necessitating specific implementation strategies:

- **Zero-Inflation**: In retail, it's common for only a subset of existing customers to return and shop with the brand in the subsequent year. Consequently, a significant portion of the data exhibits a $0 CLV. Traditional regression models, particularly linear ones, can be adversely affected by this distribution, leading to consistent underestimations. The EM-RF model, as detailed in Section 4, adeptly addresses this by bifurcating the modeling process into churn classification and CLV regression for non-churned customers. In the ensemble setup, we adopt a similar strategy: FWLS is exclusively applied to customers predicted to return by the churn classifier, while for those predicted to churn, only the EM-RF model is employed. This methodology significantly reduced CLV underestimation in the ensemble model without augmenting computational complexity.
- **High-Leverage Data Points**: Customer transaction data frequently exhibits right-skewness, with large order frequency and AOV values, particularly if B2B accounts associated with substantial purchases are present. To prevent excessive influence of large prediction errors stemming from super purchasers during training, we employ the Huber loss in place of the mean square error (MSE).
- **Negative Predictions**: Given that FWLS is a linear model employing an identity link function, it might occasionally yield implausible negative CLV predictions, even when all target values in the training dataset are positive. To counteract this issue, we implement a log transformation on the target variable prior to training.

6 Performance evaluation

In this section, we evaluate the performance of our two newly developed and implemented approaches: EM-RF (as detailed in Section 4) and FWLS (explained in Section 5). We benchmark these against three baseline models: the conventional multi-stage Random Forest model (M-RF), the EP/NBD model, and the SQ model.

For our performance assessment, we chose twelve retail brands for which CLV predictions had been active for over eighteen months. We conducted a backtesting of the model across six evaluation periods. These periods started between 2020-06-01 and 2021-09-01, and concluded between 2021-06-01 and 2022-09-01. Our results show that the enhanced multi-stage EM-RF model consistently surpasses all baselines across three key metrics: mean absolute error (MAE), total revenue percentage error (TR-PE), and Spearman's correlation coefficient. Moreover, the ensemble FWLS model improves the EM-RF's performance, particularly in terms of MAE. This improvement is consistent across all brands, irrespective of factors such as retail type, customer base size, average order value, retention rate, or the specific evaluation period.

6.1 Dataset description

Much of the existing CLV research evaluates models using one or a limited number of sample datasets over a single evaluation period, posing challenges to their generalizability. In contrast, we assess our models across twelve major consumer brands. The smallest retailer in our study contains approximately 165K customers with 389K

Table 1 Descriptive summary of twelve brands used for model evaluation. Statistics are calculated based on the latest testing period

Retailer ID	Industry	Customer Count	Trans Count	Months of Trans	AOV ($)	Retention Rate (%)
R-1	restaurant	7,456,307	19,597,605	63	144.79	41.59
R-2	cosmetics	165,665	389,459	63	78.70	33.31
R-3	apparel	10,350,599	30,772,263	79	78.98	34.59
R-4	apparel	10,028,033	41,784,788	59	108.46	42.27
R-5	apparel	11,020,490	48,872,902	59	79.37	52.88
R-6	shoes	16,784,509	51,444,085	64	155.33	29.67
R-7	apparel	6,810,236	34,034,442	68	72.09	65.87
R-8	apparel	4,764,146	25,098,505	68	177.44	55.29
R-9	sportswear	4,609,249	8,351,538	90	113.95	30.25
R-10	apparel	1,485,750	4,397,855	56	217.63	46.91
R-11	sportswear	623,286	1,922,111	68	250.60	44.15
R-12	sportswear	6,487,792	14,413,178	68	171.55	35.76

transactions, while the largest serves over 16MM customers and facilitates 51MM transactions. The data for this analysis spans from 2014 to 2022. A summary of the datasets is shown in Table 1. To maintain data privacy, each brand is represented by a Retailer ID, with brand names listed in alphabetical order.

6.2 Evaluation metrics

We adopt a standard backtesting approach with rolling windows for model evaluation. Each model is tested in six distinct periods to assess its stability over time (see Fig. 5). In practical, CLV is rarely employed as a one-time prediction. Brands monitor fluctuations in predicted CLV chronologically and frequently initiate marketing campaigns rooted in the trajectories of customers' CLV. Ensuring consistent and dependable predictive performance over extended periods is crucial for the success of such marketing strategies. To obtain a comprehensive understanding of model performance, we choose

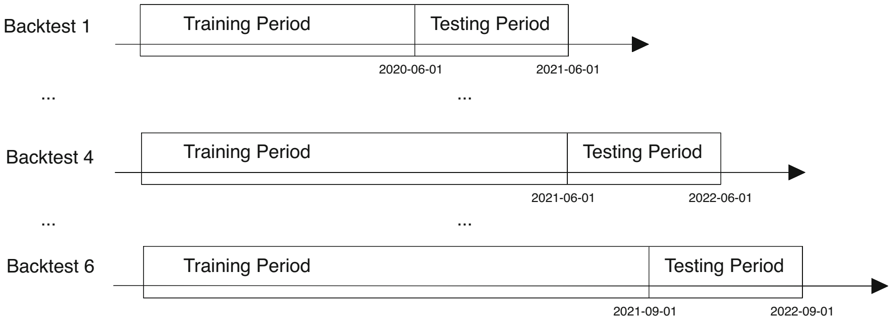


Fig. 5 Backtesting Framework

three metrics, mean absolute error (MAE), total revenue percentage error (TR-PE), and Spearman's correlation coefficient. Each metric evaluates a unique aspect of prediction quality.

For assessing prediction accuracy at the individual customer level, commonly used measures include mean squared error (MSE) and mean absolute error (MAE). We opt for MAE due to its lower sensitivity to outliers, a common occurrence in retail, where CLV distributions tend to be heavily right-skewed. MAE is defined as:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\mathrm{P}_i - \mathrm{A}_i .| \,, \tag{4}$$

where P_i and A_i are the predicted and actual CLV for customer i, respectively.

We also use TR-PE to provide an overall perspective by comparing the actual revenue with the predicted revenue across the entire customer base. It is defined as

$$\text{TR-PE} = \frac{\left|\sum \mathrm{P}_i - \sum \mathrm{A}_i\right|}{\sum \mathrm{A}_i} \cdot 100. \tag{5}$$

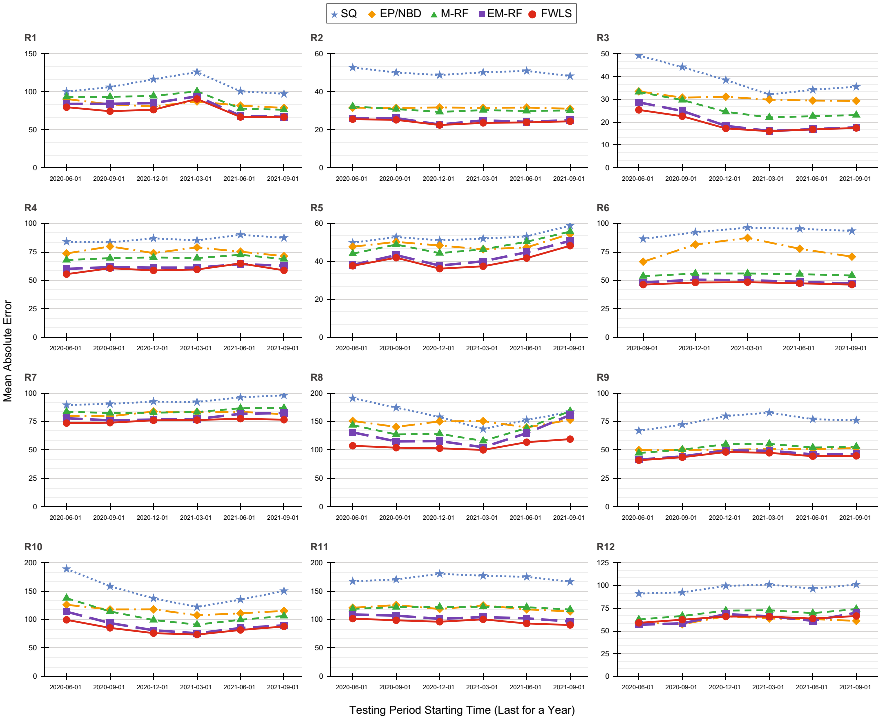


Fig. 6 Comparison of MAE between the SQ, EP/NBD, M-RF, EM-RF, and FWLS models across twelve retail brands over six evaluation periods

Brands frequently want to pinpoint their high-value customers based on relative rankings rather than absolute purchasing amount. Spearman's correlation coefficient quantifies the relationship between the ranks of actual and predicted CLVs for a specific customer, making it our third evaluation metric. Fluctuations in macro-economic conditions between training and testing phases might lead to pronounced variations in MAE and TR-RE. But Spearman's correlation coefficient generally emerges as a more resilient metric. This is based on the intuition that top-tier customers will likely continue to outspend their counterparts, even if the absolute spending amounts of all customers change.

6.3 Results

Five different models (SQ, EP/NBD, M-RF, EM-RF, and FWLS) were evaluated over six testing periods, each a quarter apart, starting from T = 2020-06-01. At T = 2020-06-01, several brands experienced severe economic turbulence due to the COVID-19 pandemic. Although our model is not designed to anticipate exogenous events, we demonstrate that our approach consistently outperforms other forecasting alternatives in the face of such challenges.

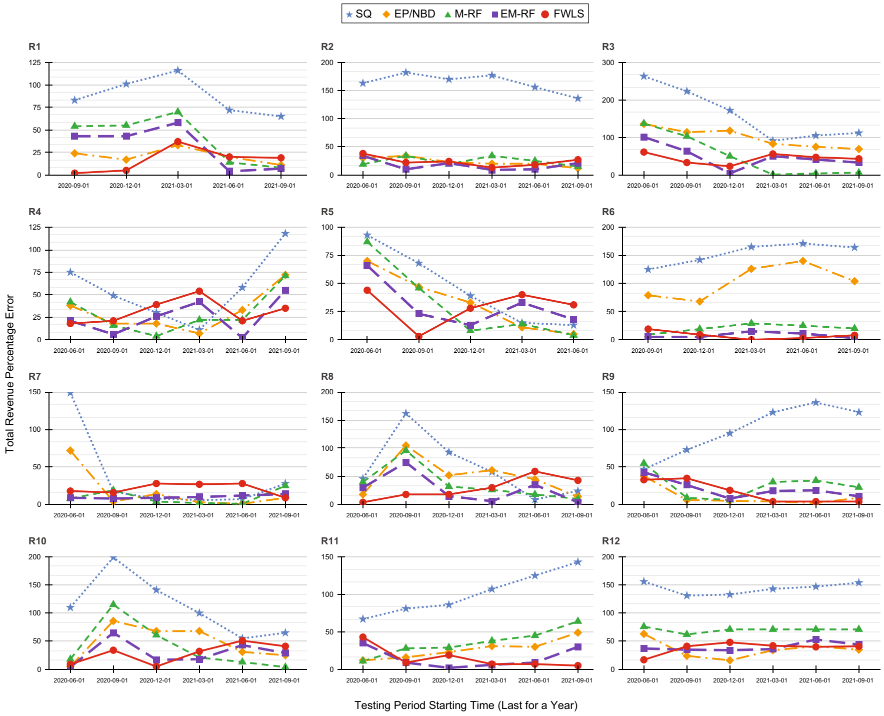


Fig. 7 Comparison of TR-PE between the SQ, EP/NBD, M-RF, EM-RF, and FWLS models across twelve retail brands over six evaluation periods

Figure 6 shows that all predictive models outperform the status-quo model nearly always, with EM-RF and FWLS consistently yielding the lowest MAE. Compared to M-RF, EM-RF reduces the MAE by an average of 8.8%, and FWLS further reduces it by 4%. Overall, the best-performing model, FWLS, outperforms the SQ model by 37.27% and the EP/NBD model by 21.57% on average. Notably, a consistent trend in model performance is observed across brands, irrespective of the testing period. Figure 7 displays the performance comparison in total revenue percentage error. While no single approach emerges as the best in every scenario, predictions using EM-RF and FWLS rarely result in large errors (> 50%). Figure 8 presents the performance comparison in terms of Spearman's correlation coefficients. M-RF, EM-RF, and FWLS perform similarly, as their predicted rankings are highly correlated, resulting in significantly higher Spearman's correlation coefficients (0.37) compared to EP/NBD (0.32) and SQ (0.27). Tables 2, 3, and 4 summarize the MAE, TR-PE, and Spearman's correlation coefficient for each model in every testing period, respectively.

We have observed that CLV modeling poses more challenges for brands with low retention rates. In such scenarios, leveraging CLV predictions can empower brands to identify and reengage churned customers. Figure 9 illustrates the relationship between a brand's retention rate and each predictive model's MAE performance improvement over the status-quo model. Generally, the predictive models, especially the two RF-

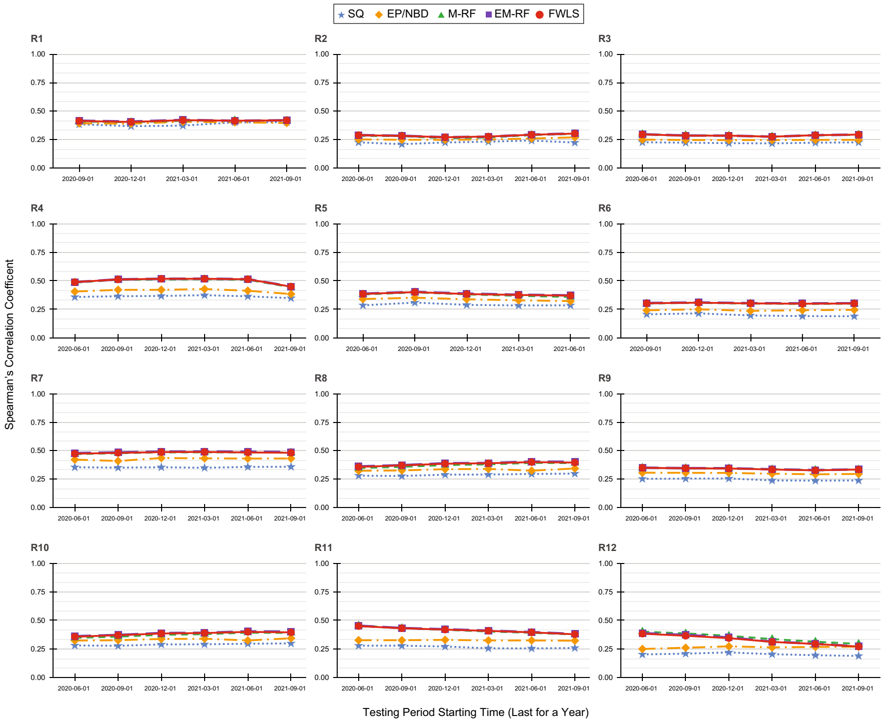


Fig. 8 Comparison of Spearman's correlation coefficients between the SQ, EP/NBD, M-RF, EM-RF, and FWLS models across twelve retail brands over six evaluation periods

Table 2 Mean absolute error of CLV predictions using five different models, evaluated in six testing periods, averaged across the twelve retail brands

Mean Absolute Error (Avg. of 12 Brands)					
Testing Period	SQ	EP/NBD	M-RF	EM-RF	FWLS
Jun ’20 – Jun ’21	88.53	66.31	69.46	61.58	58.20
Sep ’20 – Sep ’21	60.11	48.72	44.43	38.04	36.55
Dec ’20 – Dec ’21	64.46	56.99	50.19	44.50	42.50
Mar ’21 – Mar ’22	121.53	107.45	99.51	91.64	82.82
Jun ’21 – Jun ’22	124.08	91.27	88.22	75.74	71.55
Sep ’21 – Sep ’22	122.54	80.06	86.52	75.60	73.54
Avg. of All Periods	98.13	76.97	74.14	65.51	61.55
Improvement over SQ	-	21.57%	24.45%	33.25%	37.27%

Table 3 Total revenue percentage error of CLV predictions using five different models, evaluated in six testing periods, averaged across the twelve retail brands

Total Revenue Percentage Error (Avg. of 12 Brands)					
Testing Period	SQ	EP/NBD	M-RF	EM-RF	FWLS
Jun ’20 – Jun ’21	112.11	24.33	34.67	27.44	18.89
Sep ’20 – Sep ’21	118.40	55.53	36.80	32.27	33.93
Dec ’20 – Dec ’21	74.78	49.33	24.11	20.56	22.11
Mar ’21 – Mar ’22	65.93	45.33	25.40	18.93	22.80
Jun ’21 – Jun ’22	104.07	26.07	29.20	20.67	21.07
Sep ’21 – Sep ’22	142.89	35.33	67.22	37.00	31.22
Avg. of All Periods	101.31	40.07	34.79	25.60	25.24
Improvement over SQ	-	60.45%	65.66%	74.73 %	75.09%

Table 4 Spearman’s correlation coefficient of CLV predictions using five different models, evaluated in six testing periods, averaged across the twelve retail brands

Spearman’s correlation coefficient (Avg. of 12 Brands)					
Testing Period	SQ	EP/NBD	M-RF	EM-RF	FWLS
Jun ’20 – Jun ’21	0.337	0.349	0.367	0.372	0.371
Sep ’20 – Sep ’21	0.278	0.316	0.371	0.374	0.374
Dec ’20 – Dec ’21	0.269	0.310	0.357	0.364	0.361
Mar ’21 – Mar ’22	0.296	0.353	0.402	0.411	0.407
Jun ’21 – Jun ’22	0.256	0.304	0.352	0.356	0.354
Sep ’21 – Sep ’22	0.221	0.284	0.364	0.354	0.351
Avg. of All Periods	0.276	0.321	0.370	0.374	0.372
Improvement over SQ	-	16.02%	34.07%	35.35%	34.55%

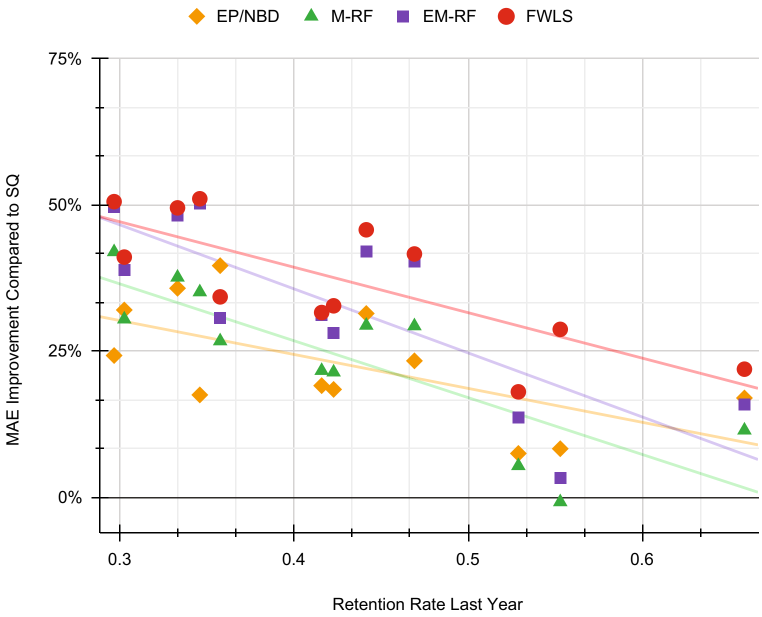


Fig. 9 The relationship between MAE and a brand's retention rate from the latest evaluation period

based ones (M-RF and EM-RF), offer greater improvement when a brand's retention rate is lower. We didn't observe strong correlations in other aspects of the data.

7 Conclusions

Our modeling approach focuses on three main strategies to consistently achieve state-of-the-art predictive performance over time across a diverse set of retail brands: 1) feature engineering that leverages high-cardinality categorical data such as demographics and product purchase history using encoders and embeddings; 2) an enhanced multi-stage CLV model designed to improve the interlinked modeling challenges of customer churn and CLV. In our backtesting results, this model consistently outperforms both a widely-used statistical model and a traditional Random Forest model, especially for brands with high price points and low retention rates; 3) feature-weighted linear stacking, which ensembles discriminative and generative models at the individual level, leading to further reduction in mean absolute error. Each of these methods has been tested, deployed, and is continuously monitored within a general ML system that was developed with two primary design considerations: 1) an infrastructure that abstracts input data, model settings, and computational resources through a configuration layer; 2) and longitudinal stability monitoring to ensure the consistent delivery of high-quality predictions and to maintain customer trust.

Acknowledgements This work is the result of dedicated team effort, made possible by the support of Amperity's Engineering and Product teams. Special acknowledgment goes to Cary Lee, Ethan Cohen, Joe

Christianson, Joyce Gordon, James Wu, Pranav Behari Lal, Stephen Meyles, and Yongsuhn Jo, for their invaluable contributions to the construction of the predictive modeling pipeline and the preparation of this paper. We are also grateful to our customers for selecting Amperity as their CDP provider, presenting us with a unique opportunity to address this real-world modeling challenge.

Data Availability Statement The data that support the findings of this study are proprietary and are not publicly available due to commercial confidentiality and privacy restrictions on handling customer data. However, summarized or aggregated results are presented within the manuscript. For questions related to the data, please contact the corresponding author.

Declarations

Disclosure of Potential Conflicts of Interest The research presented in this manuscript was fully funded by Amperity (https://amperity.com). Yan Yan and Nicholas Resnick are employees of Amperity. The authors declare that aside from this funding source, there are no potential conflicts of interest to disclose.

References

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). A novel cost-sensitive framework for customer churn predictive modeling. *Decision Analytics, 2*(1), 5. https://doi.org/10.1186/s40165-015-0014-6

Bauer, J., & Jannach, D. (2021). Improved customer lifetime value prediction with sequence-to-sequence learning and feature-based models. *15*(5). https://doi.org/10.1145/3441444

Chamberlain, B. P., Cardoso, A., Liu, C. H. B., Pagliari, R., & Deisenroth, M. P. (2017). Customer lifetime value prediction using embeddings. *KDD*

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Desirena, G., Diaz, A., Desirena, J., Moreno, I., & Garcia, D. (2019). Maximizing customer lifetime value using stacked neural networks: An insurance industry application. In *2019 18th IEEE International conference on machine learning and applications (ICMLA)* (pp. 541–544). IEEE

Fader, P., & Hardie, B. (2005). A note on deriving the pareto/nbd model and related expressions.

Fader, P. S., & Hardie, B. G. (2013). The gamma-gamma model of monetary value. *February, 2*, 1–9.

Fader, P., Hardie, B., & Lee, K. (2005). Counting your customers the easy way: An alternative to the pareto/nbd model. *Marketing Science, 24*, 275–284. https://doi.org/10.1287/mksc.1040.0098

Glady, N., Baesens, B., & Croux, C. (2009). A modified pareto/nbd approach for predicting customer lifetime value. *Expert Systems with Applications, 36*(2), 2062–2071. https://doi.org/10.1016/j.eswa.2007.12.049

Gnanadesikan, R. (2011). Methods for statistical data analysis of multivariate observations. John Wiley & Sons, ???

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., & Sriram, S. (2006). Modeling customer lifetime value. *Journal of Service Research - J SERV RES, 9*, 139–155. https://doi.org/10.1177/1094670506293810

Herbert, R. (1956). An empirical bayes approach to statistics. In *Proceedings of the third berkeley symposium on mathematical statistics and probability, vol. 1* (pp. 157–163)

Jašek, P., Vraná, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2018). Modeling and application of customer lifetime value in online retail. *Informatics, 5*. https://doi.org/10.3390/informatics5010002

Jašek, P., Vraná, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2019). Comparative analysis of selected probabilistic customer lifetime value models in online shopping. *Journal of Business Economics and Management, 20*, 398–423. https://doi.org/10.3846/jbem.2019.9597

Johannemann, J., Hadad, V., Athey, S., & Wager, S. (2019). Sufficient representations for categorical variables. arXiv:1908.09874

Johnson, S. C. (1967). Hierarchical clustering. *Psychometrika, 32*(3), 241–254.

Kalm, K., Scully, R., Haghighi, A., & Fagan, H. (2019). A report on the error rates of customer misidentification and why this is so bad for your business.

Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M. (2018). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research, 281*. https://doi.org/10.1016/j.ejor.2018.04.034

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter, 3*(1), 27–32.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013) Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119)

Platt, J., et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers, 10*(3), 61–74.

Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting your customers: Who are they and what will they do next? *Management Science, 33*(1), 1–24. https://doi.org/10.1287/mnsc.33.1.1

Sill, J., Takács, G., Mackey, L.W., & Lin, D. (2009). Feature-weighted linear stacking. *CoRR*. arXiv:0911.0460

Thai-Nghe, N., Gantner, Z., & Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. (pp. 1–8). https://doi.org/10.1109/IJCNN.2010.5596486

Vanderveld, A., Pandey, A., Han, A., & Parekh, R. (2016). An engagement-based customer lifetime value system for e-commerce. 293–302. https://doi.org/10.1145/2939672.2939693

Venkatesan, R., & Kumar, V. (2004). A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing, 68*(4), 106–125.

Wang, X., Liu, T., & Miao, J. (2019). A deep probabilistic model for customer lifetime value prediction. arXiv:1912.07753 [stat.AP]

Yan, Y., Meyles, S., Haghighi, A., & Suciu, D. (2020). Entity matching in the wild: A consistent and versatile framework to unify data in industrial applications. In: *Proceedings of the 2020 ACM SIGMOD international conference on management of data* (pp. 2287–2301)

Zhang, J., Dixit, A., & Friedmann, R. (2010). Customer loyalty and lifetime value: An empirical investigation of consumer packaged goods. *The Journal of Marketing Theory and Practice, 18*, 127–140. https://doi.org/10.2753/MTP1069-6679180202

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.