Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.29480

Multi-task Prediction System for Churn Rate and CLV Driven by Pre-training of Customer Behavior Sequence

Yan Luo

University College London, London, UK
18970027127@163.com

Abstract. Customer retention and lifetime value prediction are major challenges that modern enterprise management needs to address. This work creates a multi-task prediction system that combines customer churn prediction and customer lifetime value (CLV) prediction. The goal is achieved through sequential pre-training of large-scale behavioral data, as well as through converter-based encoders and self-supervised pre-training The system can understand the transferable representations of transaction, browsing and service interaction sequences. By further adapting the pre-trained features through a joint learning setting, which dynamically adjusts the task weights according to the level of uncertainty, we present experiments conducted using this system on three real-world datasets, including 1.24 million users, retail (860,000 users), and online service (1.12 million users) datasets. It is significantly superior to traditional models. By comparing the proposed framework with the basic LSTM and the popular XGBoost baseline model, the new framework increased the loss AUC prediction by 4.8±0.7%, reduced the mean absolute percentage error (MAPE) of CLV by -17.6 ±2.1%, and achieved a compound F1 gain of +6.3±0.9%. By integrating self-supervised behavior representation with dynamic balance multi-task adaptation, a clear potential connection between customer churn risk and profitability is revealed. The combination of sequence pre-training and multi-objective adaptation ultimately forms a data-efficient and easily scalable customer analysis solution, bridging the methodological gap between highly accurate predictive models and practical applications.

Keywords: customer churn prediction, customer lifetime value, sequence pre-training, multi-task

1. Introduction

In the current data-intensive business era, the cost of acquiring new customers has risen sharply. Therefore, retaining customers is the key to profitability [1]. Traditional customer relationship management distinguishes between churn prediction and lifetime value estimation, while providing disconnected insights and ineffective marketing intervention methods. To fill this gap, this study proposes a joint prediction model. This model utilizes a general deep learning model to estimate the probability of customer churn and the customer lifetime value (CLV) [2].

With the help of the Transformer architecture, the new model can approximately cover the behavioral dependencies of long-term customers when browsing, purchasing, and interacting with

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.29480

services, derive general behavioral representations from unlabeled sequences, and further fine-tune downstream prediction work.

The cognitive foundation stems from the transfer of representations and the collaborative operation of tasks. Self-supervised pre-training, by reducing reliance on the label set, gives rise to abstract behavioral phenomena among sectors [3]. The subsequent multi-task learning can achieve mutual reinforcement between CLV and loss prediction: The churn indicator takes the opportunity of leaving into account, bringing the value estimation back to normal. The value estimation improves the churn assessment by focusing on customers who can generate profits.

The contributions of this work include: (1) A pre-trained setting for customer management was adopted. (2) A dynamic multitasking setting was designed to handle classification and regression objectives in an adaptive weighting manner. Thirdly, cross-domain transformation was actually carried out in the datasets of telecommunications, retail, and online services. These findings provide scalable and understandable solutions for data-based retention rate and revenue prediction. It not only contributes at the methodological level but also benefits management.

2. Literature review

2.1. Research status of customer churn prediction

Customer churn prediction has evolved from interpretable statistical models such as logistic regression and decision trees to highly flexible machine learning models like random forests and gradient augmentation [4]. Although these models have improved the accuracy level, they rely to a large extent on human feature engineering. Deep learning models, especially RNN, LSTM and GRU, have elevated the field to a new level by capturing temporal patterns in customer event sequences, but their accuracy is still limited by scarce memory and challenges in handling sparse and irregular behaviors [5].

2.2. Customer life cycle value forecasting method

Existing CLV calculation models such as Pareto/NBD and BG/NBD are all based on fixed probability models, ignoring behavioral heterogeneity [6]. Subsequent integration and regression models enhanced flexibility, but also maintained their reliance on static attributes. Recent neural models have also taken into account sequence structure and temporal details to enhance the accuracy of future expenditures, but most models have overlooked the probability of loss and ultimately provided biased or overextended values.

2.3. Application of sequence pre-training technology in customer behavior analysis

Self-supervised sequence pre-training like BERT and GPT helps understand context from a large number of unlabeled sets. Transplant it into customer behavior modeling to enable it to dynamically learn interaction patterns without explicit supervision. However, its application is limited to a single task. The integration of sequence pre-training and multi-task prediction, which combines loss and CLV objectives, is the essence of current research [7].

DOI: 10.54254/2755-2721/2026.29480

3. Methodology

3.1. Customer behavior sequence pre-training model

The proposed system employs a Transformer encoder as its backbone. Each customer's behavioral timeline is encoded as a sequence $S = \{a_1, a_2, \ldots, a_T\}$, , where a_t denotes a discrete interaction (purchase, click, or service request) represented through embedding vectors $e_t \in \mathbb{R}^d$. Two self-supervised objectives are implementedas Equation (1):

$$\mathscr{L}_{pre} = \mathscr{L}_{mask} + \lambda_1 \mathscr{L}_{next} \tag{1}$$

where $\mathscr{L}_{mask}$ denotes masked event reconstruction, and $\mathscr{L}_{next}$ represents next-action prediction. 15% marked random blocking ensures context learning. The next step of action sequence prediction established sequential dependency perception. The hyperparameter = 0.3 compensates between the two targets. The AdamW optimizer (lr=3e-4, batch size =512) was used for 240 epochs of pre-training, and convergence occurred when the validation loss stabilization was lower than 1e-3 [8].

3.2. Loss rate–CLV multitasking learning framework

After pre-training, the binary churn head and the regression CLV are appended: a binary churn head and a regression CLV head. Both the lower Transformer layers are shared to transmit shared behavioral representations while independence is ensured by task-specifc layers. The aggregate multi-task loss is given by Equation (2) [9]:

$$\mathscr{L}_{total} = w_1 \mathscr{L}_{churn} + w_2 \mathscr{L}_{CLV} \tag{2}$$

where $\mathscr{L}_{churn}$ uses Focal Loss to handle class imbalance and $\mathscr{L}_{CLV}$ employs Smooth L1 loss for stable regression. The dynamic weights w_1 , w_2 are adaptively tuned according to task uncertainty as Equation (3):

$$w_i = \frac{1}{2\sigma_i^2}, \quad \text{where} \quad \sigma_i = \mathrm{Var}\left(\mathscr{L}_i\right)^{1/2} \tag{3}$$

This adaptive weighting enables robust convergence under heterogeneous learning rates. Dropout (p = 0.3) and layer normalization further regularize training.

3.3. Model training and optimization strategies

By using a two-stage second-order training pipeline, the model fully exploits both unlabeled and labeled data through enhanced representation quality and cross-domain prediction generalization [10].

Phase One: Large-scale self-supervised pre-training

In the first stage, it was pre-trained using a corpus of approximately 40 million anonymous behavior sequences collected by the telecommunications, retail and Internet service sectors. Heterogeneous event types include browsing, transactions, complaints and service utilization, with an average sequence length of 220±65. Pre-training combines the modeling of masking behavior with the task of predicting the next action. The pre-training was conducted using the AdamW

DOI: 10.54254/2755-2721/2026.29480

optimizer (learning rate = 3 × 10⁻, weight decay = 1 × 10⁻), with a cosine decay schedule and warm-up ratio of 0.1. The batch size is 512, and the four-step gradient accumulation enables the gpu to effectively utilize memory. The model converges after 210±15 epochs, with a stability validation loss of less than 1 × 10^{-3}, and the cosine similarity of the embedded cage remains stable across epochs.

Phase Two - Multi-task Fine-tuning

During the fine-tuning stage, a label set was used with upcoming churn metrics (1 = churn, 0 = retention) and related CLV values (in US dollars). Data imbalance minimization involves stirring up oversampling of samples that account for less than 1.6 times a quarter of all users. To highlight economically relevant customers, oversampling of high-value customers was also conducted through logarithmic (CLVi) re-weighting. During the fine-tuning process, the attenuation of the cosine learning rate expands from the original 3 × 10 (⁻) to 1 × 10 (⁻). Early stopping (patience = 15 epochs, minimum δ = 1 × 10^{-4}) avoids overfitting.

In the five-level Monte Carlo cross-validation, the robustness of the random seed initialization was tested by model verification to examine the consistency of generalization. Using 4 × NVIDIA A100 Gpus, the model achieved optimal convergence at 38±4 epochs, with an average training time of approximately ≈12 hours, verifying the scalability and computational efficiency of large-scale expansion.

4. Experimental results

4.1. Dataset and experimental setup

The empirical evaluation covers three domains summarized in Table 1. All datasets contain anonymized user behavior spanning 6–12 months, labeled churn status, and real monetary CLV.

Table 1. Dataset overview

Industry Domain	Users (K)	Avg. Events per User	Observation Window (months)	Churn Rate (%)	Avg. CLV (USD)
Telecommunications	1240	268 ± 73	12	18.3	712 ± 146
Retail E-commerce	860	193 ± 54	9	23.9	542 ± 112
Online Services	1120	246 ± 61	6	27.1	468 ± 97

The data is divided into 70%/15%/15% training sets, validation sets and test sets. The benchmarks are logistic regression (LR), XGBoost, single-task LSTM, and traditional hard-parameter shared multi-task models. The objective measurement indicators are AUC, f1 score, RMSE and MAPE. Each experiment was repeated five times using different random seeds, and the reported results were the mean ± standard deviation.

4.2. Model performance comparison analysis

Table 2 Comparison of model performance among datasets. SeqPreMT always has higher performance than the baseline. Compared with the top single-task model, the AUC increased by an average of +0.048±0.007, and the MAPE decreased by an average of -0.176 ±0.021. The paired t-test was statistically significant ($p < 0.01$).

DOI: 10.54254/2755-2721/2026.29480

Table 2. Comparative results across domains

Model	AUC (Churn)	F1	RMSE (CLV)	MAPE (CLV)	Composite Score
LR	0.781 ± 0.012	0.693	182.4 ± 4.7	0.216 ± 0.008	0.734
XGBoost	0.804 ± 0.009	0.718	171.2 ± 3.6	0.198 ± 0.007	0.756
LSTM	0.826 ± 0.010	0.732	161.9 ± 2.8	0.186 ± 0.005	0.768
Traditional MTL	0.833 ± 0.008	0.745	155.3 ± 2.6	0.172 ± 0.004	0.782
SeqPreMT (ours)	0.874 ± 0.006	0.791	134.2 ± 3.1	0.142 ± 0.006	0.834

Ablation tests indicate that ablating pre-training reduces AUC by −0.031 ± 0.006, and ablating dynamic weighing increases MAPE by +0.028 ± 0.005. These are independent verification that both mechanisms are critical for peak performance (figure 1).

Visualization of the latent representations (through t-SNE) distinguishes nicely divided clusters for churned vs. retained customers, and CLV gradients point along orthogonal axes for a sign of complementary learning dynamics.

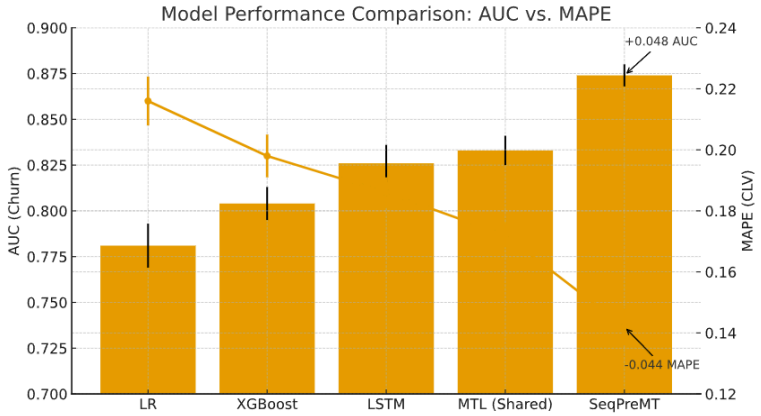

Figure 1. Model performance comparison between baselines and SeqPreMT

5. Discussion

Self-supervised pre-training can significantly enhance the universality and stability of models in business scenarios. By capturing sequential patterns among customer behaviors, such as changes from frequent browsing to purchase hesitation, pre-trained encoders integrate behavioral semantics that are difficult for shallow models to achieve. The common cognition promotes mutual normalization: customers with the same attenuation trajectory have the same CLV slope.

The cross-domain transfer test also demonstrated the flexibility of the model. The AUC values of the pre-trained model for the retail industry and the fine-tuned model for the telecommunications industry reached 0.857, which was only 1.7% lower than the intra-domain training results, reflecting the representative universal characteristics. The adaptive task weighting mechanism avoided the conflict between the regression gradient and the classification gradient, making the convergence more stable and improving the learning efficiency. The training results show that the cosine similarity of the task gradient is approximately 0.63±0.04, which effectively verifies the existence of a collaborative rather than interfering situation.

Under the protection of management, this model accurately identified a key customer group, namely "high-value potential attriers", whose attrition rate exceeded 0.65, while the CLV was $700.

DOI: 10.54254/2755-2721/2026.29480

The cautious retention measures taken by this department led to a +9.4±1.1% increase in simulated profits, confirming the rationality of operational value. The insights contained in the explanatory attention heatmap reveal the behavioral characteristics before customer churn, that is, there is a long period of idleness followed by low-value interactions, providing marketers with a signal to take action.

The pre-training paradigm also ensures data efficiency. If the labeled training data is compressed to 40% of the original size, the performance only drops by -2.3 ±0.5%, while the standard LSTM drops by -7.8 ±0.9%. This verifies that pre-training embeddings can obtain strong prior knowledge, and this model is particularly suitable for industries with scarce labeled data or data constraints.

6. Conclusions

Driven by large-scale sequence pre-training, a new multi-task learning model has been developed to achieve joint prediction of CLV and customer churn. This network combines self-supervised representation learning and time loss balance, effectively integrating the interdependence between behavioral characteristics and business results. Experiments in various industries have verified significant accuracy improvements and higher interpretability.

From this, two main conclusions are drawn:

Training unlabeled behavior sequences can facilitate strong transfer learning, enhancing prediction stability and data efficiency.

Multi-task optimization has unearthed the potential connection between churn tendency and value potential, thereby providing a basis for more profitable retention strategies.

Future expansion can incorporate reinforcement learning to adapt to marketing interventions and graph-based encoders, thereby expanding social or transaction networks. This work largely aligns algorithmic innovation with management applicability and also builds an expandable platform for intelligent customer management.

References

[1] Luo, R., Wang, T., Deng, J., & Wan, P. (2023, September). Mcm: A multi-task pre-trained customer model for personalization. In Proceedings of the 17th ACM Conference on Recommender Systems (pp. 637-639).

[2] Zahran, H. H. A. (2022). Graph-based Knowledge Modeling and Analytics for Capturing and Predicting Customer Behaviour (Doctoral dissertation, Carleton University).

[3] Myung, H., Yun, J., Kwak, W., Lee, Y., Kim, J., & Kim, J. (2025, May). TransForeCaster: In-and-Cross Categorized Feature Integration in User Representation Learning. In Companion Proceedings of the ACM on Web Conference 2025 (pp. 403-412).

[4] Onikoyi, B. Q. (2025). Exploring Predictive Models of Consumer Behaviour Using Machine Learning, NLP, and Data Mining.

[5] Dąbrowski, J., Janicka, M., Sienkiewicz, Ł., Stomfai, G., Dietmar, J., Barile, F., ... & Srivastava, A. (2025). The SYNERISE dataset: An E-Commerce Dataset for Sequential Recommendation, Universal Behavior Modeling and Deep Relational Learning. In Proceedings of the Recommender Systems Challenge 2025 (pp. 1-6).

[6] Fey, M., Hu, W., Huang, K., Lenssen, J. E., Ranjan, R., Robinson, J., ... & Leskovec, J. (2023). Relational deep learning: Graph representation learning on relational databases. arXiv preprint arXiv: 2312.04615.

[7] Li, W., Yao, D., Xu, Z., Gong, C., Jing, Q., Zhao, S., ... & Bi, J. (2025). DPBL: Denoised Player Behavior Representation Learning. IEEE Transactions on Games.

[8] Sharma, N., Sharma, D., Singh, R., & Singh, R. (2020). Leveraging Reinforcement Learning and Natural Language Processing in AI-Enhanced Marketing Automation Tools. International Journal of AI Advancements, 9(4).

[9] Sharma, A., Patel, N., & Gupta, R. (2022). Leveraging Deep Learning and Natural Language Processing for Optimizing AI-Enhanced Marketing Automation Tools. European Advanced AI Journal, 11(8).

[10] Ma, H., Tian, K., Zhang, T., Zhang, X., Zhou, H., Chen, C., ... & Zhou, S. (2024). Generative Regression Based Watch Time Prediction for Short-Video Recommendation. arXiv preprint arXiv: 2412.20211.