MEAN: Multi-Expert Adaptive Network For Customer Lifetime Value Prediction

Kelin Liu[1,2], Yao Zhou[1] (✉), Bin Liu[2], Hanjing Su[1], and Shouzhi Chen[1]

[1] WeChat Pay Research and Development Department, Tencent, China
{clingliu, yaozzhou, justinsu, easychen}@tencent.com
[2] Key Laboratory of Data Engineering and Visual Computing, Chongqing University of Posts and Telecommunications, China
Cling798as@gamil.com, liubin@cqupt.edu.cn

Abstract. Customer Lifetime Value (CLTV) is a crucial metric for evaluating the economic value that users bring to a business over their entire service cycle. Accurately predicting CLTV is essential for resource optimization, improving user retention, and maximizing return on investment (ROI). However, predicting CLTV remains challenging due to the inherent sparsity and long-tail distribution of customer spending behavior, particularly in payment scenarios where user decisions are highly dynamic and influenced by external factors. Existing methods attempt to alleviate these issues but struggle with embedding quality and distribution selection, limiting their effectiveness in capturing complex user behaviors. To address these challenges, we propose the Multi-Expert Adaptive Network (MEAN), a novel CLTV prediction framework that improves embedding representations and mitigates distribution-related errors. MEAN integrates a Multi-View Feature Express (MVFE) module to optimize multi-view representations through expert-driven feature extraction and a Distribution Adaptive Module (DAM) for soft distribution assignment, preventing error amplification from incorrect sub-distribution choice. Furthermore, we introduce an alignment mechanism to synergize MVFE and DAM via bi-directional probability alignment. Extensive offline experiments and real-world online A/B testing on the WeChat financial experimental platform demonstrate the effectiveness of MEAN.

Keywords: Customer Lifetime Value Prediction · Multi-View Features · Distribution Adaptation · Attention Joint Alignment

1 Introduction

The Customer Lifetime Value (CLTV) represents the total economic benefit that a single user brings to a product or application over their lifetime. As a core operational metric, assisting service providers in carrying out targeted marketing to improve customer retention and reduce churn rates [1, 13, 17, 6]. Therefore, accurate CLTV predictions can effectively enhance resource utilization, such as advertising costs for user acquisition and personalized service costs. This enables

the allocation of limited resources among different users to maximize the return on investment (ROI) [12].

Due to variations in individual behaviors and the inherent characteristics of customer activity over time, CLTV typically exhibits a long-tail distribution. Notably, the user attrition rate reaches approximately 90% after initial use, i.e., major users do not contribute any revenue after their registration (CLVT=0), while there are only about 10% of users transition to effective usage within one-month post-activation ((CLVT>0)).

To handle the sparsity of non-zero samples in CLTV prediction, previous works adopt a two-stage cascading architecture to divide users into several groups and train different models for each group to predict purchase propensity and potential monetary value [15, 4]. These methods typically employ ensemble machine learning models, such as random forests [15] and the XGBoost [4], which require substantial storage to maintain multiple models and struggle to capture high-level feature representations. In addtion, the two-stage cascading process can introduce error accumulation, further affecting prediction accuracy.

In recent years, end-to-end CLTV prediction models have seen significant advancements. Wang et al. [16] propose the ZILN loss function, which enables multi-objective optimization by combining purchase probability with log-normal distribution parameters. Li et al. [8] introduce a multi-expert strategy that decomposes the skewed distribution into sub-distributions, but this hard partitioning is susceptible to data noise. OptDist [18] adaptively learns optimal sub-distribution segments, but its hard distribution selection heavily depends on the accuracy of the sub-distribution assignment. MDAN [9] alleviates data sparsity through a channel weighting mechanism and a distance similarity loss to constrain the hidden-space distribution. While it employs soft distribution assignment to mitigate error propagation from sub-distribution selection, its predictions remain sensitive to scale transformations of CLTV values, particularly in scenarios with scarce positive samples. Therefore, existing methods can not properly handle the long-tail CLTV distribution.

Predicting CLTV in payment scenarios presents additional challenges due to the inherent complexity of user consumption behavior and the influence of external disturbances Consumption decisions are primarily driven by subjective awareness, characterized by high sparsity and dynamic instability [19]. For example, WeChat Credit Pay, a consumer credit product that allows users to make purchases utilizing their predetermined credit limits, has observed that users predominantly opt for this payment method in limited consumption scenarios. Moreover, exogenous variables such as promotional activities and scene adaptability interact non-linearly with users' implicit preferences, making it difficult to extract high-level features and accurately capture their latent representations.

To tackle these challenges, we propose the Multi-Expert Adaptive Network (MEAN), a novel framework for CLTV prediction. MEAN effectively mitigates the limitations of insufficient embedding representation caused by inadequate adaptation to payment scenarios. Additionally, it addresses the issues of error propagation in hard distribution selection and the high dependency on labels

in soft distribution combinations, particularly under atypical long-tail distributions. The core of MEAN is a framework based on a multi-view expert network, where Multi-View Feature Express (MVFE) module jointly optimizes multi-view features and distributions to extract complementary and robust knowledge. Specifically, instead of using the same data samples for all distributions, we pre-set multiple experts to focus on and extract different prior features of the distributions. We notice that there is a clear ordinal relationship between the distributions. Therefore, we use linear attention [20] to amplify the distinctions between the distributions, to obtain high-quality embedding representations, thereby reducing the overall complexity of CLTV modeling. To mitigate bias in MVFE, we introduce a Distribution Adaptive Module (DAM). This differs from existing methods that use the hard distribution selection criteria, as DAM can approximate the output distribution to assist MVFE, thereby preventing error amplification caused by selecting the wrong distribution. However, due to the differences in the outputs of these two modules, integrating MVFE and DAM within this framework still presents challenges. Therefore, we propose a novel alignment mechanism to address this issue, which bi-directionally aligns the probabilities output of DAM with the attention scores from MVFE. It can incorporate the distributional knowledge from DAM into MVFE, achieving a soft combination of distributions without relying on the label scale transformation. We conduct offline and online experiments on a real CLTV dataset constructed based on real users from the WeChat Payment Center, and the empirical results demonstrate the effectiveness of the proposed MEAN. The main contributions are summarized as follows:

- We propose a novel end-to-end CLTV prediction framework, MEAN, which effectively addresses the complexity of CLTV prediction and enhances adaptability to payment scenarios by optimizing high-quality embeddings for multiple candidate probability distributions.
- We design two key modules: MVFE for efficient feature encoding and DAM for distribution approximation representation. To improve model synergy, we introduce a dual-module joint optimization strategy that incorporates an attention score alignment constraint within DAM.
- Extensive offline experiments demonstrate the effectiveness of our approach, while online A/B testing on the WeChat financial experimental platform further validates the utility of MEAN in real-world marketing activities.

2 Related Work

Customer Lifetime Value modeling estimates the future revenue that new customers are expected to generate based on information about existing customers. Segmenting customers based on their CLTV and employing different marketing strategies for each segment is the initial demand of CLTV estimation. Early CLTV prediction methods focus on building rule-based or probabilistic models based on customers' historical behavior. Pareto/NBD [14] models the future purchase frequency based on customer behaviors through random process modeling,

and it is typically used in scenarios where customers can make purchases at any time. Fader and others, based on the hypothesis that users with recent purchases or relatively high purchase frequency are more likely to make future purchases, use the RFM framework [5] to group users according to the recency, frequency, and monetary value of their purchases, to estimate the CLTV of user segments. Pfeifer et at. [11] constructs a transition probability matrix using Markov chains and estimates CLTV by combining it with the initial value distribution. Machine learning methods have been widely used to directly estimate CLTV based on user features. Vanderveld et al. proposed a two-stage modeling approach [15], constructing two random forests based on user characteristics to separately predict the probability and amount of user consumption. User embedding representations are constructed using Word2Vec [3] to predict CLTV.

In recent years, end-to-end models have emerged. For example, Wang et al. designed a representative loss function, *ZILN* [16], based on the data distribution, assuming that the payment amounts follow a log-normal distribution. It uses a multi-task approach to simultaneously optimize purchase propensity, distribution mean, and distribution standard deviation. The final prediction of CLTV is the expected value from the log-normal distribution. Li et al. [8] focus on different lifecycle sequential dependencies and design the *ODMN*. They address distribution imbalance by designing the *MDME* module, which uses the divide-and-conquer approach to partition the imbalanced distribution into multiple relatively balanced sub-distributions. This module selects the appropriate expert to predict CLTV values within specific ranges. However, this approach heavily relies on the selection of sub-distributions, and modeling these sub-distributions remains challenging due to data noise, imbalance, and other factors. To address label imbalance and sparsity issues,*MDAN* [9] uses a channel learning controller and a multi-channel network to mitigate data imbalance through weighted adjustments, and designs a distance similarity loss to directly bring the hidden vectors closer to the CLTV value distribution. Unfortunately, it heavily relies on the scaling of CLTV values, which can easily lead to predicted values lacking clear distinction. *OptDist* [18] explores multiple candidate probability distributions and selects the optimal sub-distribution for each example, thereby addressing the complex and variable nature of customer lifetime value distributions. However, the use of hard selection during the inference process limits the model's adaptability.

3 Proposed Method

In this section, we introduce a novel CLTV prediction model, Multi-Expert Adaptive Network(MEAN). The overall framework of our model, as shown in Figure 1. The model consists of a multi-view feature express network (MVFE) and a distribution approximation Module (DAM). The shared layer transforms the original features into dense vectors. MVFE comprises a multi-gate mixture of experts network (MMoE) [10] and an attention mechanism, to learn unique representations of specific distributions and amplify the differences between dis-

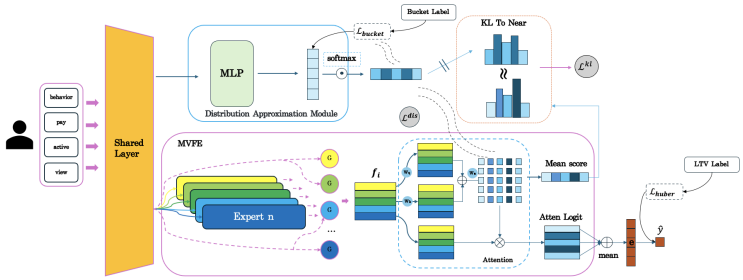

Fig. 1. The overall structure of our proposed MEAN

tributions, respectively. At the same time, our label's distribution division criterion is to set CLTV values within a certain range as a distribution, with the ranges being adjustable as needed. Zero values are considered a separate distribution. The DAM includes a distribution approximation network designed to capture the user's original distribution tendencies. Then, we describe the alignment mechanism between modules to optimize our model.

3.1 Problem Definition

Given a user group $\mathcal{U}$ and predict the total revenue that user u will bring to the product/service over some fixed lifecycle (e.g., 365 days). During this period the user's CLTV is marked as 0 if no consumption behaviour occurs, and CLTV is marked as the sum of multiple consumptions if multiple consumption behaviours occur. The training dataset $\mathcal{D} = \{(x_u, y_u) \mid u \in \mathcal{U}, y_u \in [0, +\infty)\}$ contains each sample input feature x_u and the CLTV label $y_u \geq 0$. In general, we train the model $F(\cdot)$ to predict CLTV, which can be expressed as follows:

$$\hat{y_u} = F(x_u \mid \mathcal{D}, \Theta) \tag{1}$$

where $\hat{y_u}$ is the predicted CLTV, Θ denotes the parameters of the model.

3.2 MEAN Framework

Multi-View Feature Express Module Specifically, we assume that the overall complex distribution of CLTV comprises several sub-distributions, and each user belongs to one of these sub-distributions. The diversity of input features is strongly correlated with the biased distribution of data. Noticing the varying importance of features across different CLTV distributions or segments, we observed that different users exhibit significantly different consumption behaviors based on CLTV segmentation. To address this issue, we designed a Multi-View Feature Expression network to learn and focus on the unique expressions of various distributions for different user groups. Currently, we employ M experts to represent multiple predefined distributions from different vector spaces, and use

a gating network to simulate the varying focus on features for each distribution. Based on the input feature x_u, the Multi-View Feature Express Network generates multiple distribution outputs as shown in the following formula:

$$f_i = \sum_{j=1}^{M} G\left(x_u\right)_i E_j\left(x_u\right), i \in [1, 2, ..., n] \tag{2}$$

$$G\left(x_u\right)_i = softmax\left(W_g x_u\right) \tag{3}$$

where i is the distribution sequence number, n is the total number of CLTV distributions, M is the total number of view expert networks, f_i is the embedding of the corresponding distribution output, $G\left(x_u\right)_i$ is each gating network, $W_g \in \mathbb{R}^{N \times d}$ is a trainable matrix, d is the feature dimension, and $E_j\left(x_u\right)$ is each view expert network.

Then, we use an attention mechanism to generate richer feature representations. By discarding the hard selection method of distributions, attenion-weighted aggregation can effectively improve the quality of representations, especially in scenarios with dispersed multi-distribution information. Using the attention mechanism, we allow the generation process to focus on different parts across distributions, rather than encoding the entire input into a fixed-length sequence. Importantly, we enable the model to learn to focus on what is relevant based on the existing distribution embeddings. In our setup, the input consists of the stacked representations of multiple distribution embeddings $H = concat\left[f_1, f_2, ..., f_n\right]^T$, where each distribution hidden state is denoted as h_t. There is a clear sequential order among the distributions, treating each distribution as a continuous token. We introduce an attention layer with an attention matrix $A \in \mathbb{R}^{n \times n}$, α_i is the i-th row of matrix A, where α_t is used to capture the distinctiveness of adjacent distributions. Our implementation of the attention mechanism is as follows:

$$g_t{}' = Tanh\left(W_q h_t + W_k h_t + b_{g'}\right) \tag{4}$$

$$\alpha_t = softmax\left(g_t{}' W_a + b_a\right) \tag{5}$$

where W_q and W_k represent the weight matrices for the hidden state h_t,W_a is the weight matrix for their corresponding nonlinear combination, and $b_{g'}$ and b_a are the bias terms.

Each distribution's attention-focused hidden representation is obtained by a weighted sum of the embeddings of other distributions and its current distribution's embedding similarity α_t. This is directly derived using matrix multiplication, where $A = concat\left[\alpha_1, \alpha_2, ..., \alpha_n\right]^T$. After obtaining all the distribution similarity embeddings, the final embedding e is achieved by averaging all the embeddings. This can be expressed as:

$$e = \frac{1}{n} \sum_{i=i}^{n} A_i h_i \tag{6}$$

After obtaining the final embedding of the multi-view feature representation, the CLTV prediction can be expressed as:

$$\hat{y} = FC\left(e\right) \tag{7}$$

where $\hat{y}$ is the CLTV prediction value, and FC is a fully connected layer without activation. As previously mentioned, the zero-inflated log-normal distribution [16] was proposed specifically for the CLTV distribution but is prone to extreme prediction values. However, the MSE loss is overly sensitive to these extreme values, causing the overall predictions to tend towards the mean. Therefore, we use Huber Loss to constrain $\hat{y}$ learning, enhancing discriminability between $\hat{y}$ and ensuring $\hat{y}$ remains within a controllable range. The loss function is expressed as:

$$\mathcal{L}^{cltv} = \begin{cases} \frac{1}{2}\left(y - \hat{y}\right)^2, \text{ for } |y - \hat{y}| <= \delta \\ \delta\left(|y - \hat{y}| - \frac{1}{2}\delta\right), \text{ otherwise} \end{cases} \tag{8}$$

where δ is the value that controls the turning point between the two types of loss functions.

Distribution Approximate Module Previously, we proposed a multi-view feature representation network, which aims to aggregate more influential final embeddings from the perspectives of feature and distribution debiasing. However, the specific distribution lacks clear supervisory signals. Therefore, we propose to use a bypass to construct a distribution approximation network. By using the output of such a network to constrain attention learning. The input feature x_u is projected to a high-dimensional feature vector via a simple shared layer, and then is used to produce a n dimensional feature k via MLP network:

$$k = MLP\left(ReLU\left(W_s x_u + b_s\right)\right) \tag{9}$$

Where MLP is a network structure with the last layer outputting a dimension of n, W_s and b_s are the parameters of the shared layer. We normalize the n dimensional feature vector to $\hat{b}$, ensuring the sum of all its elements to be 1, and use a multi-class classification loss to guide the learning of $\hat{b}$.The formula is as follows:

$$\hat{b_i} = \frac{e^{k_i - max(k)}}{\sum_{i=1}^{n} e^{k_i - max(k)}} \tag{10}$$

$$\mathcal{L}_{buckect} = -\frac{1}{B} \sum_{(x_u, z) \in B} \sum_{i=1}^{n} z_i log\hat{b_i} \tag{11}$$

Where B is the number of samples in a batch,z is the n dimensional one-hot vector representing the true distribution bucket label, $\hat{b_i}$ is the probability value of belonging to the $i-th$ bucket.

Constrained Attention Joint Alignment Mechanism Next, we introduce a constrained attention joint alignment mechanism in our method, Aiming to integrate the outputs of the two modules, incorporating distribution information into the MVFE module, in order to refine the final embeddings that contain distribution supervisory signals. We need to optimize the loss function with outputs from both modules. To effectively guide the direction of attention learning and prevent the attention from shifting away from focusing on effective inter-distribution information, we average the attention scores represented as $\hat{\vartheta}$. By minimizing the Kullback-Leibler (KL) divergence [2] between the normalized n dimensional feature $\hat{b}$ and $\hat{\vartheta}$, we aim to utilize the outputs of the DAM to guide the learning of attention, represented as follows:

$$\hat{\vartheta} = \frac{1}{n}\sum_{i=1}^{n} A_i \tag{12}$$

$$\mathcal{L}^{kl} = \sum_{i=1}^{n} \hat{b_i} log \frac{\hat{b_i}}{\hat{\vartheta_i}} \tag{13}$$

Where A is the attention matrix.

However, we do not want the learning of attention to be entirely guided by the output of the DAM. We also aim to transfer the information from the attention module to the DAM. Here, we use a combination of soft and hard labels through high-temperature distillation. The higher the temperature, the smoother the output probability distribution of the softmax, facilitating the transfer of knowledge from our attention module to our distribution approximation network. This achieves the purpose of mutually constrained joint alignment. The formula is expressed as follows:

$$\mathcal{L}^{dis} = \beta * \mathcal{H}\left(z, \sigma(k; T=1)\right) + \gamma * \mathcal{H}\left(\sigma(\hat{\vartheta}; T=\tau), \sigma(k; T=\tau)\right) \tag{14}$$

$$\mathcal{H}\left(p, q\right) = -\sum_{i}^{n} p_i log(q_i + \xi) \tag{15}$$

$$\sigma\left(o; T\right) = \frac{e^{o_i/T}}{\sum_{i=1}^{n} e^{o_i/T}} \tag{16}$$

where σ is the softmax function with the temperature parameter T, $\hat{\vartheta}$ is the output of the attention scores, β and γ are coefficients, ξ is a very small constant. It is worth noting that when $T = 1$, $\mathcal{H}\left(y, \sigma(k; T=1)\right)$ is equivalent to the distribution multi-classification loss $\mathcal{L}_{buckect}$. In summary, the overall loss of MEAN is defined as:

$$\mathcal{L}^{MEAN} = \mathcal{L}^{cltv} + \mathcal{L}^{kl} + \mathcal{L}^{dis} \tag{17}$$

4 Experiments

4.1 Experimental Setup

Dataset The dataset for this experiment is derived from the user growth operations of WeChat Credit Pay. WeChat Credit Pay is a consumer credit product that allows users to make purchases using their allocated credit limits. Due to customers' autonomy in choosing payment methods and amounts, along with the significant influence of concurrent marketing activities on most users, the distribution of CLTV exhibits a high level of complexity. The features of the data set include user profile data, channel-related information, and transaction records prior to the activation of the service. In addition, periodic, seasonal, and social information is also incorporated as part of the features, resulting in a total of 720-dimensional features. We sample 22 million users as the experimental dataset. In the dataset, we randomly split them into 7:1:2 as the training, validation, and test sets, respectively. Labels are defined as the total consumption amount of new users within one month, one quarter, six months, and one year after activating the WeChat Credit Pay product. Based on consumption habits and user attributes before activation by new users, we need to simultaneously estimate $cltv_{30}$,$cltv_{90}$,$cltv_{180}$,and $cltv_{365}$.

Metrics The Percentile MAPE is an evaluation metric based on Decile MAPE (DM) [16], used to evaluate the accuracy of CLTV prediction, but with finer granularity. The Gini coefficient [16] is a commonly used metric for evaluating CLTV model performance. This metric serves as a quantitative evaluation standard for the effectiveness of high-value user identification, and its value is strictly positively correlated with the model's discriminative ability: the larger the Gini coefficient, the more accurately the model distinguishes the value of top users. Spearman Correlation (SC) [16] quantitatively evaluates the ordinal consistency of the predicted values of the model, specifically representing the monotonic preservation ability of the predicted CLTV and the actual CLTV in the user value ranking. AUC is used to assess the recognition ability of high-value users. AUC focuses only on the order of relationships and can be used to evaluate the accuracy of rankings.

Baselines We compared our method with several state-of-the-art CLTV prediction methods, which are summarized as follows:

- *DNN-ZILN* [16]: A method that unifies binary classification and regression based on the log-normal distribution.
- *MTL-ZILN*: Using a multi-task learning paradigm to evaluate different CLTV periods to assist in long-term prediction.
- *ODMN* [8]: A multi-distribution multi-expert method for CLTV prediction, which divides training samples into multiple sub-distributions and buckets, estimates deviations within buckets to obtain fine-grained CLTV values.

- *MDAN* [9]: A method for predicting CLTV using multi-channel learning, where the final embedding is obtained through weighted summation fusion.
- *OptDist* [18]:An end-to-end CLTV prediction framework adaptively selects the optimal sub-distribution for each example by exploring multiple candidate probability distributions. The framework includes two modules, DLM and DSM, designed for learning sub-distributions and making distribution choices, respectively. Combined with an alignment mechanism, it enables flexible selection.

Hyperparameter Settings We use a two-layer MLP with ReLU activation functions as the shared layer. The size of the shared layer was set to [512,256], [256,128,64] for the MLP. In our main model, there is only one layer followed by a batch normalization operation. We use Adam [7] as the optimizer for our model, with a learning rate of 1e-3. The batch size is set to 1024 and the number of expert networks is set to 5. We employ an early stopping mechanism, and the model typically converges within 12 to 15 epochs. We scaled the labels and truncated them to the range [0, 20]. The parameter δ of Huber Loss and the soft label distillation coefficient γ are set to 1.0. The parameters τ of L^{dis} is set to 2.0. We also pre-divided the data into 5 distributions, with zero values being treated as a separate distribution. Our code is publicly available on an anonymous GitHub repository [3].

4.2 Performance Comparison

In Table 1, we present the evaluation results of each model on the test set. Firstly, *MTL-ZILN* outperforms *DNN-ZILN* overall because it can aggregate information from multiple periods through shared experts. Secondly, in the *ODMN* method, the error propagation caused by multiple distribution buckets can lead to significant errors in calculating the CLTV value due to misclassified distributions and buckets. In particular, We attempted to apply the multi-task prediction method from the baseline to the *MTL-MEAN*, outputting multiple periods simultaneously and summing the losses for joint optimization. Although this approach achieved some positive results, the numerous cascading losses caused the optimization direction to become unclear, resulting in performance that was not as good as single-task prediction.

The performance of our proposed MEAN model surpasses all baselines in the key metric, Percentile Mape. Moreover, it also demonstrates superior performance in both GINI and AUC metrics. This indicates that our method effectively handles imbalanced data, highlighting the robustness of our model. In addition, observing the Spearman metric, our method shows greater generalization ability in maintaining the monotonic relationship between predicted and actual values. At the same time, *MDAN* outperforms other methods in the Spearman metric for the prediction of $cltv_{180}$, indicating that the scaled values of this period label can be more easily fitted through the RankSim loss. However, this does

[3] https://anonymous.4open.science/r/ltv-F54B

Table 1. The overall performance of different models on all datasets, where the symbol ↑(↓) indicates that the higher (lower) the metric value, the better the performance.

Period	Method	Percentile Mape ↓	GINI ↑	Spearman ↑	AUC ↑
$cltv_{30}$	*DNN-ZILN*	0.7552	0.4861	0.3933	0.7295
	MTL-ZILN	0.7239	0.4960	0.4075	0.7397
	ODMN	0.6503	0.4449	0.4050	0.7395
	MDAN	0.1658	0.5080	0.4309	0.7428
	OptDist	0.3925	0.4918	0.4098	0.7418
	MTL-MEAN	0.1630	0.5112	0.4298	0.7518
	MEAN	**0.1258**	**0.5127**	**0.4358**	**0.7549**
$cltv_{90}$	*DNN-ZILN*	0.5755	0.5056	0.4206	0.7361
	MTL-ZILN	0.5269	0.5044	0.4233	0.7393
	ODMN	0.4855	0.4771	0.4215	0.7388
	MDAN	0.1332	0.5177	0.4566	0.7501
	OptDist	0.1918	0.5098	0.4322	0.7450
	MTL-MEAN	0.1033	0.5192	0.4596	0.7536
	MEAN	**0.0753**	**0.5275**	**0.4687**	**0.7571**
$cltv_{180}$	*DNN-ZILN*	0.5262	0.5081	0.4231	0.7300
	MTL-ZILN	0.4551	0.5085	0.4292	0.7359
	ODMN	0.3591	0.5068	0.4382	0.7417
	MDAN	0.1034	0.5239	**0.4846**	0.7487
	OptDist	0.1222	0.5118	0.4562	0.7469
	MTL-MEAN	0.1129	0.5244	0.4669	0.7588
	MEAN	**0.0927**	**0.5298**	0.4805	**0.7615**
$cltv_{365}$	*DNN-ZILN*	0.3821	0.4968	0.4272	0.7249
	MTL-ZILN	0.2890	0.5069	0.4376	0.7313
	ODMN	0.1779	0.4988	0.4257	0.7247
	MDAN	0.1206	0.5232	0.4829	0.7587
	OptDist	0.1063	0.5072	0.4804	0.7529
	MTL-MEAN	0.0972	0.5194	0.4920	0.7597
	MEAN	**0.0871**	**0.5292**	**0.4987**	**0.7649**

not imply generalization capability. This further validates the effectiveness and generality of our carefully designed model in predicting CLTV.

4.3 Ablation Study

In this section, we conduct ablation experiments to evaluate the effectiveness of each innovative module of *MEAN*. We compare the differences in Percentile Mape and AUC for the four periods of the overall sample across different modules of *MEAN*. Percentile Mape and AUC are the primary reference metrics for distinguishing the capabilities of our model. They represent the accuracy of the model in predicting CLTV and the precision in identifying top users, respectively. Main content: (1) Without $\mathcal{L}^{kl}$: Remove the KL divergence loss term from the alignment mechanism; (2) Without the soft label distillation term $\mathcal{L}_2^{dis}$: Remove the soft label distillation term from the alignment mechanism; (3) Without DAM and $\mathcal{L}^{dis}$ and $\mathcal{L}^{kl}$: Remove the entire alignment mechanism, only using the Multi-View Feature Express Network module.

Table 2. Percentile Mape results of MEAN and its variants for different prediction periods.

Method	$cltv_{30}$	$cltv_{90}$	$cltv_{180}$	$cltv_{365}$
MEAN	**0.1258**	**0.0753**	**0.0927**	**0.0871**
$Without - \mathcal{L}^{kl}$	0.1918	0.1706	0.1162	0.0953
$Without - \mathcal{L}_2^{dis}$	0.1727	0.1001	0.1228	0.1121
MVFE	0.2317	0.1571	0.1129	0.1279

Table 3. AUC of MEAN and its derivative methods for different prediction periods on the dataset.

Method	$cltv_{30}$	$cltv_{90}$	$cltv_{180}$	$cltv_{365}$
MEAN	**0.7549**	**0.7571**	**0.7615**	**0.7649**
$Without - \mathcal{L}^{kl}$	0.7514	0.7532	0.7521	0.7588
$Without - \mathcal{L}_2^{dis}$	0.7523	0.7516	0.7500	0.7558
MVFE	0.7539	0.7531	0.7508	0.7551

Then we summary the results of the ablation experiments in Table 2 and Table 3. It demonstrates that our alignment mechanism can effectively improve the accuracy and stability of CLTV prediction. When the soft label distillation term and KL divergence loss term are not used, the overall performance of CLTV prediction decreases, indicating that the model's spontaneous attention focus cannot be controlled. We further investigated the impact of the two constraint terms within the alignment mechanism. When only one constraint term is used, the short-term prediction performance without the KL divergence loss is better than that without the soft label distillation term, while the long-term prediction performance shows the opposite trend. This indicates that the two constraint terms focus on different aspects, potentially leading to excessive guidance in attention learning. This validates our design of the constrained attention joint alignment mechanism. Specifically, *MVFE* enables us to obtain better embedded representations of features, though it has certain instability. By employing the constrained attention joint alignment mechanism, the overall performance of the model is improved, resulting in highly accurate CLTV predictions over both long and short periods.

4.4 Online A/B Test

To further validate the effectiveness of MEAN on real-world applications, we perform an A / B test on the financial experiment platform of WeChat Pay. Our experiment divided users into two groups, those who have activated the feature and those who have not, to ensure homogeneity among users. We assigned 50% of the traffic to the control group and 50% to the experimental group. The model was employed to estimate users' $cltv_{365}$, and marketing efforts were directed

towards users with high $cltv_{365}$. To facilitate the observation of experimental results, the modeling label in the $cltv_{365}$ estimation was set as the user's total loan amount one year later. The experimental group's strategy was to target the top 10% as predicted by the model, while the control group's strategy was to randomly target 10%. For each marketing campaign, we separately observed the online effects after 7 days, 14 days, and 30 days. We then calculated the activation rate (only for users who had not activated), the loan rate, and the average daily loan amount during the experiment period.

Table 4. The improvement of various metrics in A/B testing

Method	Activation Rate	Loan Rate	Averge Daily Loan Amounts
Not Activated UPLIFT - 7	1.51%	0.08%	2.20%
Not Activated UPLIFT - 14	2.47%	0.11%	4.31%
Not Activated UPLIFT - 30	4.24%	0.18%	7.01%
Activated UPLIFT - 7	-	0.03%	1.98%
Activated UPLIFT - 14	-	0.06%	2.77%
Activated UPLIFT - 30	-	0.11%	5.56%

To ensure company privacy, Table 4 only presents the improvement values of our method relative to the control group. Our method demonstrates superior performance, and due to the large base of traffic, the increase in revenue is also highly significant. This further validates the effectiveness of our proposed model and the accuracy in identifying high-value users.

4.5 Hyperparameter Analysis

In this section, we investigate the impact of two key hyperparameters on our method: the parameter δ of Huber Loss and the number of sub-distributions n. We primarily focus on the evaluation of Percentile MAPE for the overall sample in practical business scenarios, as this metric reflects the model's prediction accuracy. Additionally, we will use the predictions to guide the allocation of marketing resources. Therefore, we will discuss the impact of different parameters on the performance of *MEAN* with respect to this metric. The parameter δ controls the transition of the CLTV loss. Figure 2 shows the performance of the framework under different values of δ. As δ decreases, L^{cltv} increases, resulting in a smaller alignment mechanism loss initially, which weakens the constraints. Furthermore, We change the number of sub-distributions in the set $\{2, 3, 4, 5, 6, 7\}$. Figure 3 shows the model performance under different numbers of sub-distributions. Similarly, as the number of sub-distributions is increasing, the overall loss of the alignment mechanism is increasing, shifting the focus of the overall optimization of the framework, leading to a decrease in prediction accuracy. In real-world scenarios, we recommend practitioners search these hyperparameters according to the key metrics in the corresponding applications.

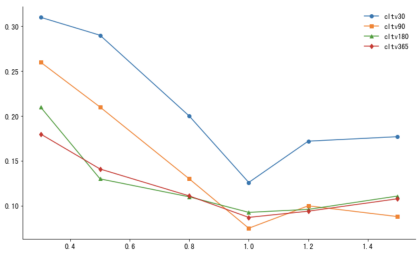

Fig. 2. The impact of parameter δ on model performance under different prediction periods.

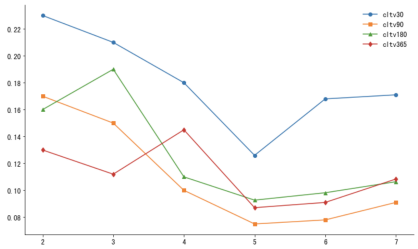

Fig. 3. The impact of the number of sub-distributions on model performance under different prediction periods.

5 Conclusion

In this paper, we propose a new framework for customer lifetime value prediction called MEAN. MEAN obtains feature perspective expressions through multiple expert networks for different CLTV distributions and uses an attention mechanism to amplify the differences between distributions. The aggregated user embeddings contain feature diversity and distributional distinctiveness. Additionally, we propose a joint alignment mechanism that uses DAM to approximate the distribution of the original features, constraining the direction of attention. At the same time, the attention scores guide the output of DAM, achieving mutual constraints, thereby making the optimization more effective. In this way, MEAN pays more attention to the differences between different features and different distributions, making the CLTV prediction capability more intuitive. Finally, our method has achieved considerable gains in both offline experiments and online applications on real-world industrial datasets, with consistent results demonstrating the effectiveness of MEAN.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (62302074) and the Science and Technology Research Program of Chongqing Municipal Education Commission (KJQN202300631).

References

1. Paul D Berger and Nada I Nasr. Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*, 12(1):17–30, 1998.
2. Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
3. Pei Pei Chen, Anna Guitart, Ana Fernández del Río, and Africa Periánez. Customer lifetime value in video games using deep learning and parametric models. In *2018 IEEE international conference on big data (big data)*, pages 2134–2140. IEEE, 2018.

4. Anders Drachen, Mari Pastor, Aron Liu, Dylan Jack Fontaine, Yuan Chang, Julian Runge, Rafet Sifa, and Diego Klabjan. To be or not to be... social: Incorporating simple social features in mobile game customer lifetime value predictions. In *proceedings of the australasian computer science week multiconference*, pages 1–10, 2018.
5. Peter S Fader, Bruce GS Hardie, and Ka Lok Lee. Rfm and clv: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, 42(4):415–430, 2005.
6. Bowei He, Yunpeng Weng, Xing Tang, Ziqiang Cui, Zexu Sun, Liang Chen, Xiuqiang He, and Chen Ma. Rankability-enhanced revenue uplift modeling framework for online marketing. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5093–5104, 2024.
7. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
8. Kunpeng Li, Guangcui Shao, Naijun Yang, Xiao Fang, and Yang Song. Billion-user customer lifetime value prediction: an industrial-scale solution from kuaishou. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 3243–3251, 2022.
9. Wenshuang Liu, Guoqiang Xu, Bada Ye, Xinji Luo, Yancheng He, and Cunxiang Yin. Mdan: Multi-distribution adaptive networks for ltv prediction. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 409–420. Springer, 2024.
10. Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1930–1939, 2018.
11. Phillip E Pfeifer and Robert L Carraway. Modeling customer relationships as markov chains. *Journal of Interactive Marketing*, 14(2):43–55, 2000.
12. Patricia Pulliam Phillips. *Return on investment (ROI) basics*. Association for Talent Development, 2023.
13. Ziv Pollak. Predicting customer lifetime values–ecommerce use case. *arXiv preprint arXiv:2102.05771*, 2021.
14. David C Schmittlein, Donald G Morrison, and Richard Colombo. Counting your customers: Who-are they and what will they do next? *Management Science*, 33(1):1–24, 1987.
15. Ali Vanderveld, Addhyan Pandey, Angela Han, and Rajesh Parekh. An engagement-based customer lifetime value system for e-commerce. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 293–302, 2016.
16. Xiaojing Wang, Tianqi Liu, and Jingang Miao. A deep probabilistic model for customer lifetime value prediction. *arXiv preprint arXiv:1912.07753*, 2019.
17. Yunpeng Weng, Xing Tang, Liang Chen, Dugang Liu, and Xiuqiang He. Expected transaction value optimization for precise marketing in fintech platforms. *arXiv preprint arXiv:2401.01525*, 2024.
18. Yunpeng Weng, Xing Tang, Zhenhao Xu, Fuyuan Lyu, Dugang Liu, Zexu Sun, and Xiuqiang He. Optdist: Learning optimal distribution for customer lifetime value prediction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2523–2533, 2024.
19. Mingzhe Xing, Shuqing Bian, Wayne Xin Zhao, Zhen Xiao, Xinji Luo, Cunxiang Yin, Jing Cai, and Yancheng He. Learning reliable user representations from

volatile and sparse data to accurately predict customer lifetime value. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3806–3816, 2021.

20. Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. Opentag: Open attribute value extraction from product profiles. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1049–1058, 2018.