

Collaborative-Enhanced Prediction of Spending on Newly Downloaded Mobile Games under Consumption Uncertainty

Peijie Sun
Quan Cheng Laboratory
DCST, Tsinghua University
Beijing, China
sun.hfut@gmail.com

Yifan Wang
DCST, Tsinghua University
Beijing, China
yf-wang21@mails.tsinghua.edu.cn

Min Zhang*
DCST, Tsinghua University
Beijing, China
z-m@tsinghua.edu.cn

Chuhan Wu
Noah's Ark Lab, Huawei
Beijing, China
wuchuhan15@gmail.com

Yan Fang
DCST, Tsinghua University
Beijing, China
fangy21@mails.tsinghua.edu.cn

Hong Zhu
Consumer Cloud Service Interactive
Media BU, Huawei
Shenzhen, China
zhuhong8@huawei.com

Yuan Fang
Consumer Cloud Service Interactive
Media BU, Huawei
Shenzhen, China
frank.fy@huawei.com

Meng Wang
School of Computer Science and
Information Engineering, Hefei
University of Technology
Hefei, China
eric.mengwang@gmail.com

ABSTRACT

With the surge in mobile gaming, accurately predicting user spending on newly downloaded games has become paramount for maximizing revenue. However, the inherently unpredictable nature of user behavior poses significant challenges in this endeavor. To address this, we propose a robust model training and evaluation framework aimed at standardizing spending data to mitigate label variance and extremes, ensuring stability in the modeling process. Within this framework, we introduce a collaborative-enhanced model designed to predict user game spending without relying on user IDs, thus ensuring user privacy and enabling seamless online training. Our model adopts a unique approach by separately representing user preferences and game features before merging them as input to the spending prediction module. Through rigorous experimentation, our approach demonstrates notable improvements over production models, achieving a remarkable **17.11%** enhancement on offline data and an impressive **50.65%** boost in an online A/B test. In summary, our contributions underscore the importance of stable model training frameworks and the efficacy of collaborative-enhanced models in predicting user spending behavior in mobile

gaming. The code associated with this paper has also been released at the following link¹.

CCS CONCEPTS

• **Information systems** → **Computational advertising**; **Collaborative filtering**; **Data stream mining**.

KEYWORDS

User Spending Prediction, Collaborative-enhanced Model, Revenue Optimization

ACM Reference Format:

Peijie Sun, Yifan Wang, Min Zhang, Chuhan Wu, Yan Fang, Hong Zhu, Yuan Fang, and Meng Wang. 2024. Collaborative-Enhanced Prediction of Spending on Newly Downloaded Mobile Games under Consumption Uncertainty. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion)*, May 13–17, 2024, Singapore, Singapore. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3589335.3648297>

1 INTRODUCTION

Online mobile games have become a part of many people's daily lives. Different game developers strive to earn money from mobile game players. However, convincing players to spend money depends not solely on the quality of the designed in-game goods. The game store's recommendation strategy is even more crucial. One potent way to achieve higher total revenue is by inferring how much the users may pay for newly downloaded games and formulating tailored recommendation strategies, such as the targeted Return On Ad Spend (tROAS) [43]. The task of inferring how much users pay for newly downloaded games can be treated as predicting the LifeTime Value (LTV) of these users [43]. The task of predicting LTV has been a topic of study for many years [10, 21, 23, 39, 45].

*Corresponding author.

This work is supported by the Natural Science Foundation of China (Grant No.U21B2026, 72188101), the fellowship of China Postdoctoral Science Foundation (No.2022TQ0178), and Huawei (Huawei Innovation Research Program). We also thank MindSpore for the partial support of this work, which is a new deep learning computing framework.



This work is licensed under a Creative Commons Attribution International 4.0 License.

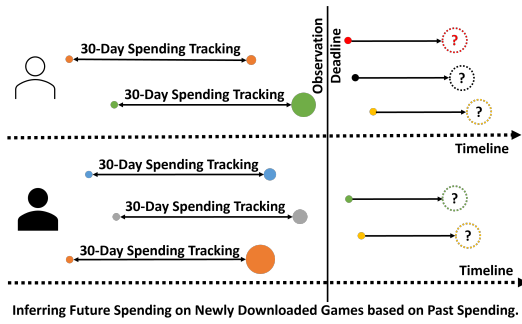


Figure 1: This illustration showcases our collected data. It includes observed spending behavior by users leading up to an observation deadline. Users’ spending habits are tracked from the day they download the game, with a 30-day observation period to gauge their spending during the subsequent 30 days. Different colored circles represent different games, with larger circles indicating higher spending. Using this observed data, we aim to predict users’ future spending within the next T days for newly downloaded games.

However, our approach represents a departure from these work, which studied how to predict a user’s future expenditure based on past expenditures. Instead, our focus is on predicting the spending money of users for any new game they download. This method is a novel approach to maximizing revenue for game developers.

One approach to this task is to convert LTV prediction into conversion rate(CVR) prediction, a technique employed by industry giants like Google, Tencent, and NetEase [43, 43, 52, 55]. Based on this approach, we need to prepare two things: data collection and model design. Given the continuous spending behavior of users and the varying life cycles of users in games, collecting the total amount spent by users is impractical. Instead, we use a 30-day observation window and consider the total spending amount within this period on newly downloaded games as the user’s LTV. In other words, the LTV would be treated as the label to train the model which is designed for the CVR task. By minimizing the LTV label and the prediction of the CVR model, we can treat the final learned CVR model as an LTV prediction model. The illustration of the dataset we collected can refer to Figure 1. Our current production environment utilizes a classical cross-feature CVR model as the backbone model. This model inputs the users’ historical behaviors, such as downloaded and paid game apps, and games’ profiles like game ID and other related metadata. The model aims to predict the LTV of the user-game pair, and we optimize it by minimizing the difference between predicted spending and real observed spending.

Upon reviewing the spending prediction task, we argue that using only the cross-feature model as the CVR model is somewhat not enough, as this type of model does not explicitly consider the collaborative relationship [16, 17, 24, 26, 29, 36, 37, 49] between users. However, collaborative signals are very important for predicting spending amounts. Because a user is more likely to spend more money on her satisfied product. Utilizing the collaborative signals can help the model better find the product that meets the target users’ preferences. However, giving collaborative capability to the

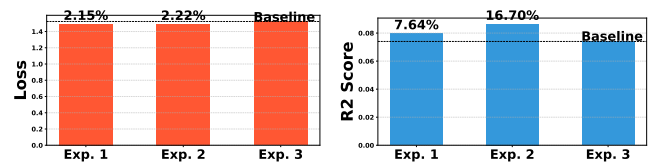


Figure 2: Repeated three production experiments, labeled Exp. 1, Exp. 2, and Exp. 3, were conducted using a private dataset to evaluate the spending money prediction capacity. RMSE-based loss is used as objective loss, and R2 Score [1] is used as model effectiveness evaluation. The worst-performing model is marked with the baseline. The numbers above each bar indicate improvements compared to the baseline. For example, the Exp. 1 bar, represents a 2.15% improvement in training loss and a 7.64% improvement in the R2 Score over the worst-performing experiment.

current cross-feature production model is not a simple task. This is mainly due to the unique characteristics of our dataset. Our data is characterized by many zero values, high label variance in non-zero values, and extremely high values, which brings significant challenges for consistently training and evaluating the effectiveness of the designed model.

The dataset’s characteristics are summarized in Table 1. To illustrate, the dataset contains many zero values, but this doesn’t mean users always spend nothing on games. Inaccurate consumption tags can harm model optimization and performance. High label variance and extreme values disrupt optimization and training stability. Relying solely on labels for model accuracy is problematic due to unpredictable user spending patterns. As a result, we risk creating a model effective only on the validation dataset, lacking robustness for broader use. Though data characteristics also affect existing production models, experiments show that different models can perform well under the production strategy despite fluctuations. We suppose that the possible reason is adequate data can help mitigate uncertainties from the collected dataset. Nevertheless, operating within the production environment continues to pose challenges for exploring new design models. For instance, in Figure 2, three experiments were conducted with the production model, revealing that while training loss remains stable, the overall performance, as measured by the R2 Score metric, varies significantly. Compared to the worst experiment, training loss only improves by 2.15% and 2.22%, while the R2 Score shows substantial improvements of 7.64% and 16.70%.

Our study has two main goals. First, we propose a framework to enhance model training and evaluation stability by addressing data characteristics. Second, we investigate collaborative signal modeling within this framework to develop a collaborative-enhanced solution. In our framework, we adhere to two key principles for effective model deployment. Firstly, we use spending data as labels during model training, employing a regression loss function. Secondly, we prioritize model stability and evaluation certainty by standardizing user spending mitigating label variance and extreme values. Compared with the value-based metric (minimizing the difference between the predicted and real spending money), we utilize the leave-one-out ranking metric [17] to evaluate the

trained model. When evaluating the model’s effectiveness, we only evaluate whether the model can predict which game a user downloaded relatively accurately, without evaluating whether the user pays, and how much, because user downloading behavior is more easily learned in the dataset than the uncertain paying behavior. We also name this task the downloaded game prediction task. The experimental results also show that our proposed framework can bring out the capabilities of different models and that models validated as effective in our framework are also considered effective in predicting the amount of money a user will spend in the production environment, i.e., in the spending money prediction task.

While designing the collaborative-enhanced model, we address two key questions: First, how can we model the collaborative signal without using user IDs, which are typically essential for collaborative filtering, in order to protect user privacy? To tackle this, we represent user preferences using their historical game app downloads. We employ a multi-layer perceptron (MLP) with download history as input, obtaining the collaborative representation by linearly combining user preference and game data. Second, how can we integrate this collaborative signal into the existing production model? We use the same behavior data for both the deployed cross-feature model and collaborative signal modeling. Since the online model doesn’t explicitly capture the collaborative signal, we propose enhancing spending prediction by concatenating the collaborative feature with the cross-feature model output and inputting these concatenated features into the subsequent spending prediction module. With no need for user IDs during model training and shared historical download game IDs across data samples, the model can be updated based on streaming data, making it adaptable for online training environments. Moreover, this approach ensures user privacy protection during model training.

In summary, the contributions of our work can be summarized as: 1. Given the uncertainty surrounding users’ consumption habits in terms of time and money, we have developed a framework designed to expedite exploring the collaborative-enhanced model. This framework is precisely engineered to ensure stable model training and evaluation processes.

2. We have introduced a collaborative-enhanced model that prioritizes user privacy by not relying on user IDs. This approach allows for model training within a data streaming environment, as user IDs are unnecessary.

3. Our collaborative-enhanced model has proven effective in two settings: First, on the offline business dataset, it outperformed the production model by **17.11%** in the 30-day spending money prediction task. Second, in the online A/B test, it achieved a **50.65%** improvement in user payment revenue over two weeks compared to the production model.

2 DATASET STATISTICS

Our research aims to understand how to predict users’ future spending on newly downloaded games. To collect the dataset, we adopted a straightforward approach. We initiated a T-day observation period for each user and their newly downloaded game. During this time, we closely monitored the user’s spending on the game within the T-day window since the game’s download date. The data collection spanned from August 1, 2022, to August 31, 2022. Let’s take

Table 1: Statistics Of The Collected Private Dataset.

Consumption		Download Game Apps in Two Weeks	
Min Cost	0.01	Min Length	1
Max Cost	421,387	Max Length	10
Avg Cost	227.15622	Avg Length	3.78
Std	2130.0593	Std	3.41
Medium	11.0	Medium	2
# of Non-zero	623,075	# of All Download Apps	7,507
# of Zero	29,076,679	# of Filtered Paid Games	432

August 7, 2022, as an example to provide a clearer illustration of the collected data. On this date, we gathered information regarding how much users spent on games they had downloaded on July 7, 2022. This extended observation period was essential to accumulate a sufficient volume of user behavior data for our analysis. The statistics of the collected data can be referred to Table 1. Given our primary focus on paid games, we implemented a filter to exclude games with spending below a certain threshold. We cannot disclose the specific threshold used to safeguard company proprietary information. This filtering process accounts for the disparity between the total number of downloaded games and the number of paid games. The downloaded game applications are crucial for modeling user preferences in our subsequent work.

From the collected data, we have the following observations: **Prevalence of Zero Values:** The dataset prominently displays a substantial prevalence of zero values, encompassing an impressive count of 29,076,679 zero samples, in stark contrast to the 623,075 non-zero samples. This notable contrast implies a significant presence of users who did not partake in any spending activity. It’s important to note that due to the limited observation time window, these zero values do not necessarily indicate that users consistently spent nothing on the corresponding games. **High Variance Among Non-Zero Samples:** While the average consumption cost (Avg Cost) appears relatively low at 227.16, the standard deviation (Std) stands at a remarkably high 2130.06. This disparity implies substantial variability within non-zero consumption costs, with some values significantly deviating from the average. The resulting high standard deviation underscores the substantial variance in the dataset. **Extreme Values:** Notably, the maximum consumption cost (Max Cost) reaches an extreme value of 421,387, indicating the existence of outliers with exceptionally high spending. Similarly, the elevated standard deviation (Std) indicates significant data points deviating from the mean. These findings collectively highlight the presence of outliers, potentially attributed to users with exceptionally high expenditures.

Directly using these significantly high-variance consumption amounts as labels would introduce varying degrees of gradient values to the trainable parameters, potentially severely disrupting their normal learning process. Additionally, during the model evaluation phase, predicting user behaviors concerning these zero and extreme values could be hindered by the uncertainty surrounding user spending time and amounts. This uncertainty might result in models selected based on a specific validation set lacking the required level of generalization.

Table 2: Notation Table.

$\mathcal{U} = (u_1, u_2, \dots, u_{ \mathcal{U} })$	User set
$\mathcal{P} = (p_1, p_2, \dots, p_{ \mathcal{P} })$	Paid Game set
$\mathcal{D} = (d_1, d_2, \dots, d_{ \mathcal{D} })$	Downloaded Game set
s_{up}	Spent money from user-game pair (u, p)
t_u^{180}	Total spending of user u in the past 180 days
f_u^{180}	Total payment times of user u in the past 180 days
h_u^d	Downloaded game apps list for the user u

3 PROPOSED FRAMEWORK AND COLLABORATIVE-ENHANCED MODEL

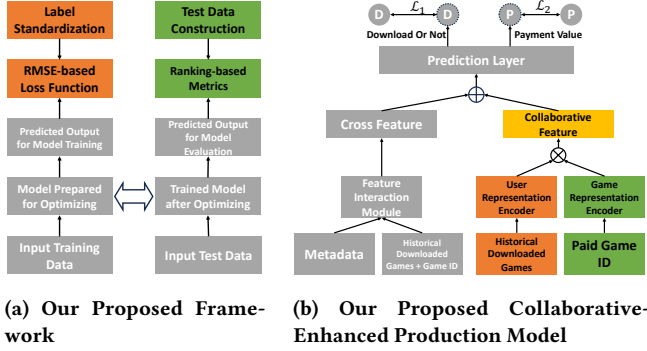


Figure 3: The Comprehensive Architecture of Our Proposed Framework and Model. The modules colored other than gray represent the new modules we have introduced.

In this section, we will first introduce the proposed framework to accelerate the exploration of newly designed models and the collaborative-enhanced model we have developed. Initially, we will introduce the notations employed in our proposed model, which can be found in Table 2.

3.1 Proposed Framework

Our proposed framework is designed to address the unique characteristics of the collected datasets, including numerous zero values, high variance in non-zero labels, and exceptionally high values. We have enhanced both the training and evaluation strategies within the production environment. The advanced modules we have introduced are depicted as non-gray modules in Figure 2(a). In the left section of Figure 2(a), we have introduced label standardization and an RMSE-based loss function to train the newly designed model. Label standardization mitigates the adverse effects of high label variance in non-zero and highly high values during model optimization. The RMSE-based loss function assesses whether the newly designed model can capture patterns under the same labels and training conditions as those in the production environment. In the right section of Figure 2(a), we employ a ranking-based method to evaluate the model’s performance. This method verifies whether the proposed model can predict users’ downloaded games from a mix of positive and negative examples. Such information is obtainable from the training data and helps reduce the impact of uncertain user spending patterns and detailed monetary data.

3.1.1 Label Standardization. To standardize the spending labels, we aim to uncover their underlying information. We argue that the absolute amount spent does not necessarily reflect a user’s preference for a game. Some low-cost games may also hold value for users. User spending habits and cumulative spending for the game can influence larger or smaller spending amounts. The standardized label serves to assess the game’s value to the user. Treating this value as the label is reasonable, as it helps the model identify valuable games for the user. This information is crucial for predicting game downloads and estimating potential spending, as users tend to spend more on games they find valuable.

Specifically, we collect the users’ total spending t_u^{180} and payment times f_u^{180} in the past 180 days before the observation time and cumulative spending for the games \mathcal{S}_p to standardize the spending money labels. Given user u and paid game p , we utilize the following equation to standardize the collected spending money s_{up} :

$$\tilde{s}_{up} = f(t_u^{180}, f_u^{180}, \mathcal{S}_p, s_{up}). \quad (1)$$

The function f denotes the standardization function, and the notation \tilde{s}_{up} denotes the standardization cost money.

We have adopted a strategy that involves first calculating the standardization values separately from the game’s and user’s perspectives and then combining them to form the final standardization value. For game-sided standardization, we compute the corresponding standardized spending value as follows:

$$\tilde{s}_p = \frac{s_{up} - AVG(\mathcal{S}_p)}{STD(\mathcal{S}_p)}, \quad (2)$$

Where $AVG(\cdot)$ represents the average function, used to calculate the mean of all historical spending for game p , and $STD(\cdot)$ represents the standard deviation of all historical spending for game p .

From the user’s perspective, we employ the following equation to calculate the user-sided standardization value:

$$\tilde{s}_u = \frac{s_{up}}{t_u^{180} / f_u^{180}}. \quad (3)$$

During the standardization process, our goal is to rescale the consumption data of different users to a similar scale to eliminate scale differences among them.

Due to the distinct scales between the values obtained by users and the standardized label values obtained by games, we combine all the standardized labels of users and games to normalize their respective standardized labels. Subsequently, we sum and aggregate them to obtain the final standardized labels, represented as $\tilde{s}_{up} = 0.5\tilde{s}_p + 0.5\tilde{s}_u$.

3.1.2 Model Optimization. In the model optimization stage, we use the MSE loss to optimize the backbone model within the framework. The objective loss can be calculated with:

$$\mathcal{L}_{MSE} = \frac{\sum_{(u,p) \in \mathcal{O}} (\tilde{s}_{up} - \hat{s}_{up})^2}{|\mathcal{O}|}, \quad (4)$$

Where \mathcal{O} denotes all observed user-game consumption records. And \hat{s}_{up} denotes the predicted output of the backbone model.

3.1.3 Model Evaluation. When assessing the effectiveness of the trained model, our objective is to determine whether the model can predict which games users might download from a mixture of positive and negative games. This task relies on information that can be captured during the training process. Specifically, for each test sample, we randomly selected 100 games from all paid games to serve as negative examples. This strategy is also employed in BPR [29] and NeuMF [17]. We evaluate the model’s performance using ranking-based metrics, including NDCG@K and HR@K [17]. HR@K indicates whether the user’s interacted game ranks in the top-K game list. On the other hand, NDCG@K assesses whether the interacted game ranks first. Generally, the higher a game is ranked, the greater the value of the corresponding metric, indicating superior performance by the trained model.

3.2 Collaborative-Enhanced Model

This section will introduce how to model the collaborative signal and enhance the production model. The updates to the proposed model can be observed in the non-gray modules of Figure 2(b). The gray section represents the simplified structure of our online production model. To safeguard the company’s proprietary information, we won’t delve into it extensively in this paper. Since the production model utilizes the same data, such as users’ historical download game apps and paid game IDs, we argue that it can capture implicit collaborative signals. Our proposed explicit collaborative signals should prove valuable in enhancing the model’s predictive power for users’ spending. As for why we don’t replace it with our proposed collaborative signal, it’s primarily because the production model employs a complex mapping function to transform features into the final spending amount. The input features should include metadata, such as users’ historical spending data. However, this paper doesn’t delve into how to combine the historical spending input with our proposed collaborative signals. We treat the collaborative signal as helpful and rely on the existing production environment to harness its value.

First, we will introduce how to model the collaborative signal without using user IDs to protect users’ privacy. The goal of user representation is to identify users with similar preferences, and for this purpose, we can leverage users’ historical behaviors to represent their embeddings. Ultimately, we choose to utilize users’ historical downloaded games as they are easier to collect and rich in information.

Given a user u ’s downloaded game list $h_d^u = \{d_1, d_2, \dots, d_{10}\}$, we can obtain the user’s preference representation as follows:

$$\mathbf{v}_u = MLP \left(\left[\mathbf{e}_1 \quad \mathbf{e}_2 \quad \dots \quad \mathbf{e}_{|h_d^u|} \right] \right), \quad (5)$$

where $\mathbf{e}_{|h_d^u|}$ denotes the embedding of the $|h_d^u|$ -th downloaded game in user u ’s download history.

The collaborative signal can be modeled using the following equation:

$$\mathbf{v}_{up} = \mathbf{v}_u * \mathbf{v}_p, \quad (6)$$

where \mathbf{v}_p represents the embedding of the paid game.

Next, we use the concatenate operation to combine the collaborative signal \mathbf{v}_{up} with the existing production model:

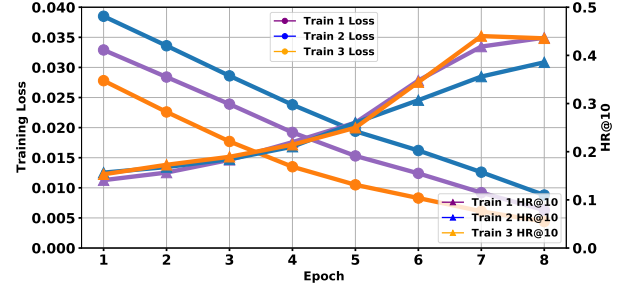


Figure 4: Training and Evaluation Process Curve

$$\mathbf{v}_{concat} = \left[\mathbf{v}_{up} \quad \mathbf{v}_{up}^{fi} \right] \quad (7)$$

The final spending amount can be predicted using the complex mapping function currently employed in the production environment:

$$\hat{s}_{up} = f(\mathbf{v}_{concat}) \quad (8)$$

4 EXPERIMENTS

In this section, we aim to validate the effectiveness of our proposed framework and collaborative-enhanced model. The experiments can be divided into two main parts. The first part is designed to confirm the viability and stability of our proposed framework in the downloaded game prediction task. We seek to verify whether our framework can aid the designed model in uncovering hidden patterns within the dataset. Additionally, we aim to assess whether models with varying capabilities can perform at their full potential. In the second part, we evaluate the effectiveness of our proposed collaborative-enhanced model in both offline and online production environments in the spending money prediction task.

4.1 Experimental Settings

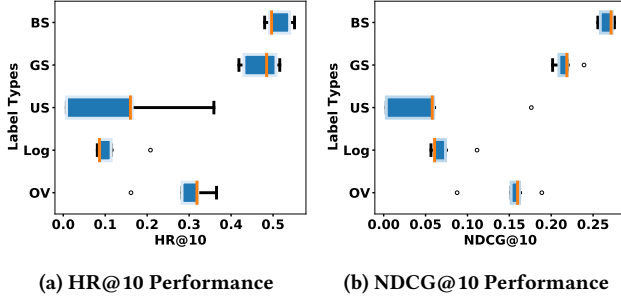
The latent factor dimension size is configured at 8, while the learning rate is set to $1e-5$. Subsequently, we employ a 4-layer MLP with layer sizes [8, 16, 32, 8] to model users’ preference representations. The Adam optimizer guides the optimization process, which updates the trainable parameters. For training, we utilize the PyTorch deep learning framework on a GPU server equipped with a 32GB GPU card, 64GB of memory, and a 32-core 2.3 GHz CPU.

4.2 Effectiveness of Our Proposed Framework

4.2.1 Framework Availability. In this study, we conducted multiple training sessions to assess the effectiveness of our recommendation framework. Our primary focus was on whether the model could uncover hidden patterns within the dataset. We closely monitored the model’s performance throughout the training process. To ensure the stability of our framework, we repeated the experiment three times and generated training and evaluation curves depicted in Figure 4. Consistently, we observed that performance metrics, such as HR@10, displayed consistent and notable improvements across all training sessions. These findings highlight the model’s capacity to enhance its recommendation capabilities as it undergoes training.

Table 3: Performance of Recommendation Models on HR@K and NDCG@K.

Method	HR@K↑			NDCG@K↑		
	K=1	K=5	K=10	K=1	K=5	K=10
MF [26]	0.0185	0.1275	0.271	0.0094	0.0687	0.1186
FM [28]	0.0415	0.1653	0.3834	0.0262	0.0893	0.1207
Wide & Deep [8]	0.0495	0.2839	0.4542	0.0306	0.1807	0.2386
DCNV2 [42]	0.0702	0.3322	0.5024	0.0403	0.1990	0.2546
OURS	0.0830	0.3687	0.5477	0.0524	0.2266	0.2841

**Figure 5: Performance of our proposed model under different label standardization types. BS denotes both-sided, GS denotes game-sided, US denotes user-sided, Log denotes logarithmic, and OV denotes origin values.**

4.2.2 Framework Stability - Different Backbone Models. In this section, to assess the stability of our proposed framework, we aim to determine its effectiveness in enabling various models to perform at their full potential. We have chosen a classical collaborative-based model, mf [26], and a classical cross-feature model, fm [28]. Additionally, we have included two competitive cross-feature models, [8] and [42]. The results can be found in Table 3. Our proposed model is denoted as **OURS**, representing an enhanced version of DCNV2 that explicitly models collaborative signals.

The results consistently revealed a linear improvement in performance from MF to **OURS**, indicating an advancement in modeling techniques. Particularly, **OURS** exhibited superior performance across all K values in both metrics, underscoring its effectiveness in prioritizing user-downloaded games within candidate game lists. An interesting trend noted was the increase in HR@K and NDCG@K scores with higher K values for all models, reflecting the broadening opportunity to capture user preferences. Additionally, the performance gap between models varied with different K values, with significant disparities becoming evident at higher Ks. These insights collectively demonstrate the robustness of 'OURS' in enhancing recommendation relevance and precision compared to other models.

Overall, the results indicate that our proposed framework assists various models in realizing their full potential and serves as inspiration. It highlights that the collaborative signal modeling method and the combination approach can be effectively applied in a production environment.

4.2.3 Framework Stability - Label Standardization. In this section, we aim to verify the effectiveness of our proposed label standardization strategies. The five label standardization methods are Original

Table 4: Evaluation of Model Generalization in Different Environments. CoV(Coefficient of Variation) [2] is a measure of data variability. CoV↓(↑): a lower(higher) value means more consistent(greater variability).

Environment	Metric	Mean	Std	CoV↓
Production	R2Score	0.02907	0.01389	0.47785
Our Framework	HR@10	0.42701	0.01942	0.04547
	NDCG@10	0.23274	0.01128	0.04848

Values(OV), Logarithmic(Log), User-sided(US), Game-sided(GS), and Both-sided(BS). OV denotes we treat original spending values as labels directly. Log denotes we utilize the logarithmic value of the spending values as labels. US denotes we only treat the user-sided standardization label as a supervision signal, such as Equation 3. Similarly, the GS denotes we only treat the user-sided standardization label as a supervision signal, such as Equation 2. From Figure 5, we find Original Values show a moderate, steady performance, underscoring the raw data's inherent predictive strength. Conversely, the Logarithmic approach consistently underperforms, suggesting it might not fit this data set best. While effective in certain scenarios, the user-sided method falls short in terms of its overall performance. In a comprehensive evaluation, it lags behind other approaches. The Game-sided strategy scores high across all labels, making it a generally effective option. However, the star performer is the Both-sided approach, excelling in every category. This method's success is likely due to its holistic consideration of user and game factors. This insight highlights the Both-sided method as the most effective overall. Furthermore, the compactness of the box plot corresponding to both sides indicates the stability of the results achieved through this standardized approach.

4.2.4 Framework Stability - Evaluation. This section aims to demonstrate that our proposed framework allows for a fairer evaluation of our learned model's generalization capabilities. Our framework effectively measures the model's generalization because it assesses whether our proposed model can accurately identify games users have downloaded from a mix of positive and negative games. In contrast, the production model's task is to predict users' future spending, which can be uncertain and unstable, leading to variable model performance from one day to another. To validate this, we conducted experiments using user spending data collected for training from August 1, 2022, to August 24, 2022. Subsequently, we evaluated the model's performance daily from August 25, 2022, to August 31, 2022, generating daily performance metrics. We exclusively assessed the learned model using the R2Score metric in the production environment. However, within our proposed framework, we evaluated its performance using HR@10 and NDCG@10 metrics. We used the same backbone model for both evaluation approaches to ensure a fair comparison. To simplify the presentation of results and account for metric scale differences, we provided the mean, standard deviation, and coefficient of variation (CoV) [2] for the performance metrics over the seven days. Please refer to the Table 4 for detailed results.

This table shows that the Coefficient of Variance (CoV) value within our framework is notably lower compared to the production environment. This observation supports the idea that our proposed framework can effectively identify models with stable performance

Table 5: Overall Performance on Private Dataset (Offline Testing). * indicates significant superiority over the top comparison (t-test, $p < 0.05$). $\uparrow(\downarrow)$: the larger(the lower) the better.

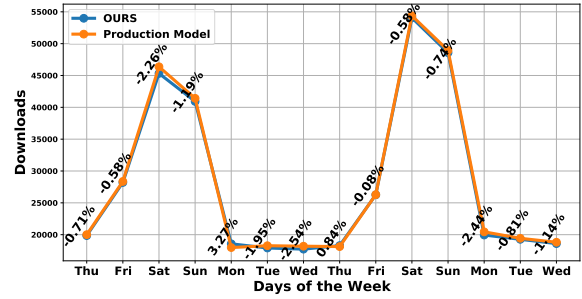
Method	All Samples			Samples with Payment >0	
	RMSE \downarrow	R2 Score \uparrow	AUC \uparrow	RMSE \downarrow	R2 Score \uparrow
Linear	247.16	0.01958	0.7517	1770.4	0.02588
MLP	248.11	0.01209	0.6549	1789.0	0.00538
Random Forest [45]	248.87	0.00599	0.7317	1803.0	0.01035
XGBoost [7]	246.40	0.02565	0.8336	1786.3	0.00835
ZILN [43]	242.56	0.05576	0.8601	1713.0	0.07721
MDME [21]	243.71	0.04395	0.8482	1743.6	0.07843
Production Model	240.16	0.07436	0.8629	1710.5	0.10291
OURS	238.50*	0.08708*	0.8693*	1698.5*	0.12844*
Improv.	0.69%	17.11%	0.74%	0.70%	24.81%

in the downloaded game recommendation task. Furthermore, models that exhibit superior performance within our proposed framework will likely excel in the user spending prediction task. Our evaluation approach remains unaffected by the uncertainty surrounding spending time and amounts. Moreover, the model’s training is primarily based on spending money. Consequently, the capabilities demonstrated by models within our framework will likely seamlessly transition into an online environment.

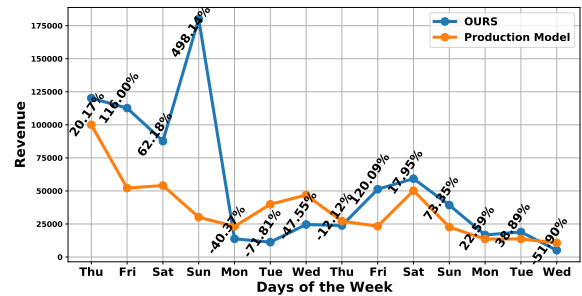
4.3 Effectiveness of Our Proposed Collaborative-Enhanced Model

4.3.1 Evaluation on Offline Business Dataset. To evaluate the effectiveness of our proposed collaborative-enhanced production model, we conducted experiments on a private dataset collected between August 1, 2022, and August 31, 2022. We divided the data into two sets: the first 30 days were used for training, and the data from the last day were reserved for testing. In addition to our production model, we compared it to other competitive models, including ZILN [43], developed by Google, and MDME [21], proposed by KuaiShou. To provide a comprehensive performance comparison, we evaluated the models on two subsets of the data: all samples and samples with spending greater than zero. In the table below, our proposed model is denoted as "OURS" in Table 5. The experiments are used to evaluate the performance of all models in predicting the spending money with RMSE and R2 Score [1]. AUC is used to evaluate whether the proposed model can predict whether the user downloaded the target game.

From Table 5, we can find our proposed model has surpassed all other models in all performance metrics, including the Linear Regression model, Multi-Layer Perceptron (MLP), Random Forest, XGBoost, Zero-Inflated Negative Binomial Regression (ZILN), Mixed Discrete and Continuous Output Model (MDME), as well as the existing production model. Our proposed model achieved the best results on all samples and those with payments greater than zero, across Root Mean Square Error (RMSE), R2 Score, and AUC. Particularly in terms of the R2 Score, it showed an improvement of 17.11% for all samples and 24.81% for samples with payments greater than zero, compared to the production model. These findings undoubtedly illustrate that the collaborative-enhanced model we have designed can significantly enhance the performance of existing production models for spending money prediction tasks.



(a) Download Data for the Date Range 20231019 to 20231101



(b) Revenue Data for the Date Range 20231019 to 20231101

Figure 6: User Game Downloads and Revenue on Newly Downloaded Games Comparison of Our Proposed Model and Production Model in an Online A/B Test from 2023-10-19 to 2023-11-01. The values corresponding to all data points are relative to protect the company’s property.

4.3.2 Online A/B Test. An online A/B test was conducted from October 18, 2023 for a two-week observation period. The test was implemented in the Editor Recommendation Module on the Huawei Game Store, with 30% of the recommended games downloads and experiment group. Our proposed collaborative-enhanced model was deployed in the experiment group to predict users’ spending on newly downloaded games in the next 30 days. Based on this prediction, different games were recommended to different users. The production model served the control group. We focused on users’ recommended games downloads and in-game purchases.

Results demonstrate that while the overall download rate decreased slightly by 0.88%, the proposed model achieved a 50.65% improvement in revenue compared to the production model. The significant revenue gains, despite the minor decrease in downloads, suggest our model was able to recommend games that users were more likely to enjoy. This could reduce situations where users download more games due to unsatisfying recommendations, which explains the slight decrease in overall downloads.

Figure 6 shows more detailed download and revenue patterns. Further analysis of the A/B test data indicates that the proposed model attained substantial revenue improvements consistently across weekdays and weekends. This points to the robustness of our collaborative approach across different user activity profiles. Moreover, some weekly cycles were also observed. After obtaining confirmation from the business unit, we conducted a larger-scale

test in the relevant recommendation scenarios and ultimately confirmed the formal implementation of this improvement into the respective product. As of February 2, 2024, the latest information we have obtained indicates that our proposed algorithm has achieved an **18.43%** increase in revenue in all traffic tests in some recommendation scenarios.

5 RELATED WORK

5.1 Customer Lifetime Value Prediction

In this paper, the research problem we study focuses on predicting users' spending on newly downloaded games [46]. Forecasting spending is challenging due to uncertainties in timing and amount. However, LTV prediction has a long history of research since the 1920s. In 1920, M. Greenwood introduced RFM and NBD for disease and disaster prediction [12]. In 1959, A.S.C. Ehrenberg applied RFM and NBD ideas to marketing [9]. In 1987, D.C. Schmittlein extended NBD to Pareto/NBD by mixing NBD with Gamma distribution of exponentials [32]. Recent LTV studies can be categorized based on machine learning methods and whether a two-stage prediction is used. By ML methods, there are Bayesian approaches like NBD, Pareto/NBD, random forest-based methods [45], and neural network models such as paper [33], TSUR [52], perCLTV [55], and ExpLTV [54]. By prediction stage, two-stage methods separately predict payers and payment amounts [5, 39], or payment frequencies and amounts per payment [11, 40]. Non-two-stage methods directly predict spending from user/product features and historical data [14, 23, 30]. However, most existing works predict future spending based on past transactions, which is inapplicable to our new game download scenario. In summary, while LTV prediction has been studied for decades, forecasting user spending on newly downloaded games remains an open challenge, rendering most prior arts inapplicable. Developing suitable solutions for this problem requires exploring novel directions. We argue that one challenge of this task is the very oddly distributed user spending money label [38], which is also addressed in ZILN [43] and MDME [21]. In this paper, we argue collaborative signals can improve spending prediction, but label noise and imbalance in the dataset hinder model exploration.

5.2 Cross-Feature Models

The user spending prediction task can be treated as a conversion rate (CVR) prediction problem. Hence, we can refer to cross-feature models used in CVR prediction. These models can automatically learn complex relationships between different features, such as FM [28], Wide&Deep [8], DeepFM [13], xDeepFM [22], NFM [15], etc. Each model designs specific modules to enhance automatic feature engineering. Due to their strong modeling capabilities, DCN [41] and DCNV2 [42] have recently been considered state-of-the-art. We employ in our production environment, we employ DCN, considering its time efficiency and good performance. Some other models like AutoInt [34] and CAN [4] have also shown superior results, but their high time complexities prevent actual deployment. Cross-feature models enable efficient feature engineering by automatically capturing intricate feature interactions. The DCN family balances modeling power and time complexity, making it prevalent in real-world systems.

5.3 Collaborative-based Methods

To infer users' potential spending on newly downloaded games without historical interaction records, collaborative filtering approaches are naturally promising solutions. Existing collaborative methods can be categorized into memory-based and model-based. Classical memory-based algorithms include item-based methods [27, 31], while typical model-based approaches contain matrix factorization models like PMF [26] and BPR [29]. Some hybrid methods combine memory-based and model-based strategies, such as SVD++ [19] and NESCL [37]. With the success of deep learning, neural collaborative filtering models like NeuMF [17] have emerged. To address the sparsity issue, solutions like graph-based methods and harder negative sampling have been proposed [6, 16, 25]. Self-supervised learning has been introduced to enhance recommender model training [18, 20, 37, 49]. Other domains like social recommendation [35, 50, 51] and news recommendation [3] have also benefited from modeling collaborative signals. In addition, because the collected user interaction data contains noise and bias, some work has also been proposed to address such issues, such as [44]. However, applying existing collaborative filtering approaches to our problem still faces some challenges. First, user IDs are unavailable, so new ways to model collaborative signals without IDs need to be explored. Second, full retraining on entire datasets is infeasible, so collaborative models that allow efficient incremental updates are preferred. Lastly, production systems already model some collaborative information via cross-features. Effectively combining explicit and implicit collaborative signals is non-trivial. Although methods like Hi-Transformer [48] and NRMS [47] learn user representations without IDs, they do not consider the dynamic update and signal combination issues. While CFFNN [53] incorporates collaborative filtering and cross-features, it focuses less on integrating explicit and implicit signals.

6 CONCLUSION AND FUTURE WORK

In this work, we proposed a framework to enable stable training and evaluation of models predicting user spending on newly downloaded games. This addresses challenges from unpredictable user behavior and data characteristics like label variance and extremes. Within the framework, we developed a collaborative-enhanced model that represents user preferences and game features separately before concatenating them to predict spending. The main contributions are: 1) The framework allows robust model exploration given uncertain user spending habits by standardizing spending data. 2) The collaborative model enhances prediction accuracy without relying on user IDs, enabling online training while protecting user privacy. And 3) experiments showed the collaborative model outperforms production models, increasing prediction accuracy by **17.11%** on offline data and revenue by **50.65%** in an online A/B test. Overall, the framework and collaborative modeling approach promise to maximize revenue from mobile games through more accurate user spending prediction, while ensuring model stability and user privacy. In the future, we aim to improve game value estimation in three aspects: enhancing model generalization, testing scalability with different datasets, and exploring new applications in streaming platforms and e-commerce.

REFERENCES

- [1] 2023. https://en.wikipedia.org/wiki/Coefficient_of_determination
- [2] 2023. https://en.wikipedia.org/wiki/Coefficient_of_variation
- [3] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 336–345.
- [4] Weijie Bian, Kailun Wu, Lejian Ren, Qi Pi, Yujing Zhang, Can Xiao, Xiang-Rong Sheng, Yong-Nan Zhu, Zhangming Chan, Na Mou, et al. 2022. CAN: feature co-action network for click-through rate prediction. In *Proceedings of the fifteenth ACM international conference on web search and data mining*. 57–65.
- [5] Benjamin Paul Chamberlain, Angelo Cardoso, CH Bryan Liu, Roberto Pagliari, and Marc Peter Deisenroth. 2017. Customer lifetime value prediction using embeddings. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1753–1762.
- [6] Lei Chen, Le Wu, Richang Hong, Kun Zhang, and Meng Wang. 2020. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 27–34.
- [7] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
- [8] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishu Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.
- [9] Andrew SC Ehrenberg. 1959. The pattern of consumer purchases. *Journal of the Royal Statistical Society Series C: Applied Statistics* 8, 1 (1959), 26–41.
- [10] Peter S Fader, Bruce GS Hardie, and Ka Lok Lee. 2005. RFM and CLV: Using iso-value curves for customer base analysis. *Journal of marketing research* 42, 4 (2005), 415–430.
- [11] Peter S Fader, Bruce GS Hardie, and Jen Shang. 2010. Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science* 29, 6 (2010), 1086–1108.
- [12] Major Greenwood and G Uduy Yule. 1920. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal statistical society* 83, 2 (1920), 255–279.
- [13] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).
- [14] Sunil Gupta, Dominique Hanssens, Bruce Hardie, Wiliam Kahn, V Kumar, Nathaniel Lin, Nalini Ravishanker, and S Sriram. 2006. Modeling customer lifetime value. *Journal of service research* 9, 2 (2006), 139–155.
- [15] Xiangnan He and Tat-Seng Chua. 2017. Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 355–364.
- [16] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.
- [17] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.
- [18] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.
- [19] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 426–434.
- [20] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *Ieee Access* 8 (2020), 193907–193934.
- [21] Kumpeng Li, Guangcui Shao, Najun Yang, Xiao Fang, and Yang Song. 2022. Billion-user Customer Lifetime Value Prediction: An Industrial-scale Solution from Kuaishou. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 3243–3251.
- [22] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1754–1763.
- [23] Edward C Malthouse and Robert C Blattberg. 2005. Can we predict customer lifetime value? *Journal of interactive marketing* 19, 1 (2005), 2–16.
- [24] Kelong Mao, Jieming Zhu, Jimpeng Wang, Quanyu Dai, Zhenhua Dong, Xi Xiao, and Xiuqiang He. 2021. SimpleX: A simple and strong baseline for collaborative filtering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1243–1252.
- [25] Kelong Mao, Jieming Zhu, Xi Xiao, Biao Lu, Zhaowei Wang, and Xiuqiang He. 2021. UltraGCN: ultra simplification of graph convolutional networks for recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1253–1262.
- [26] Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. *Advances in neural information processing systems* 20 (2007).
- [27] Xia Ning and George Karypis. 2011. Slim: Sparse linear methods for top-n recommender systems. In *2011 IEEE 11th international conference on data mining*. IEEE, 497–506.
- [28] Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*. IEEE, 995–1000.
- [29] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).
- [30] Remi M Sakia. 1992. The Box-Cox transformation technique: a review. *Journal of the Royal Statistical Society Series D: The Statistician* 41, 2 (1992), 169–178.
- [31] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.
- [32] David C Schmittlein, Donald G Morrison, and Richard Colombo. 1987. Counting your customers: Who-are they and what will they do next? *Management science* 33, 1 (1987), 1–24.
- [33] Rafet Sifa, Julian Runge, Christian Bauckhage, and Daniel Klapper. 2018. Customer lifetime value prediction in non-contractual freemium settings: Chasing high-value users using deep neural networks and SMOTE. (2018).
- [34] Weiping Song, Chence Shi, Zhiping Xiao, Zhijian Duan, Yewen Xu, Ming Zhang, and Jian Tang. 2019. AutoInt: Automatic feature interaction learning via self-attentive neural networks. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1161–1170.
- [35] Peijie Sun, Le Wu, and Meng Wang. 2018. Attentive recurrent social recommendation. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 185–194.
- [36] Peijie Sun, Le Wu, Kun Zhang, Xiangzhi Chen, and Meng Wang. 2023. Neighborhood-Enhanced Supervised Contrastive Learning for Collaborative Filtering. *IEEE Transactions on Knowledge and Data Engineering* (2023).
- [37] Peijie Sun, Le Wu, Kun Zhang, Xiangzhi Chen, and Meng Wang. 2023. Neighborhood-Enhanced Supervised Contrastive Learning for Collaborative Filtering. *IEEE Transactions on Knowledge and Data Engineering* (2023).
- [38] Arturo Valdivia. 2021. Customer lifetime value in mobile games: a note on stylized facts and statistical challenges. In *2021 IEEE Conference on Games (CoG)*. IEEE, 1–5.
- [39] Ali Vanderveld, Addhyan Pandey, Angela Han, and Rajesh Parekh. 2016. 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*. 293–302.
- [40] Rajkumar Venkatesan and Vita Kumar. 2004. A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of marketing* 68, 4 (2004), 106–125.
- [41] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*. 1–7.
- [42] Ruoxi Wang, Rakesh Shivanna, Derek Cheng, Sagar Jain, Dong Lin, Lichan Hong, and Ed Chi. 2021. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the web conference 2021*. 1785–1797.
- [43] Xiaojing Wang, Tianqi Liu, and Jingang Miao. 2019. A deep probabilistic model for customer lifetime value prediction. *arXiv preprint arXiv:1912.07753* (2019).
- [44] Yifan Wang, Peijie Sun, Min Zhang, Qinglin Jia, Jingjie Li, and Shaoping Ma. 2023. Unbiased Delayed Feedback Label Correction for Conversion Rate Prediction. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2456–2466.
- [45] Than Than Win and Khin Sundee Bo. 2020. Predicting customer class using customer lifetime value with random forest algorithm. In *2020 International Conference on Advanced Information Technologies (ICAIT)*. IEEE, 236–241.
- [46] Chuhan Wu, Qinglin Jia, Zhenhua Dong, and Ruiming Tang. 2023. Customer Lifetime Value Prediction: Towards the Paradigm Shift of Recommender System Objectives. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1293–1294.
- [47] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 6389–6394.
- [48] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Hi-Transformer: hierarchical interactive transformer for efficient and effective long document modeling. *arXiv preprint arXiv:2106.01040* (2021).
- [49] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development*

- in information retrieval*. 726–735.
- [50] Le Wu, Junwei Li, Peijie Sun, Richang Hong, Yong Ge, and Meng Wang. 2020. Diffnet++: A neural influence and interest diffusion network for social recommendation. *IEEE Transactions on Knowledge and Data Engineering* 34, 10 (2020), 4753–4766.
- [51] Le Wu, Peijie Sun, Yanjie Fu, Richang Hong, Xiting Wang, and Meng Wang. 2019. A neural influence diffusion model for social recommendation. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 235–244.
- [52] Mingzhe Xing, Shuqing Bian, Wayne Xin Zhao, Zhen Xiao, Xinji Luo, Cunxiang Yin, Jing Cai, and Yancheng He. 2021. 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*. 3806–3816.
- [53] Ruiyun Yu, Dezhi Ye, Zhihong Wang, Biyun Zhang, Ann Move Oguti, Jie Li, Bo Jin, and Fadi Kurdahi. 2021. CFFNN: Cross feature fusion neural network for collaborative filtering. *IEEE Transactions on Knowledge and Data Engineering* 34, 10 (2021), 4650–4662.
- [54] Shijie Zhang, Xin Yan, Xuejiao Yang, Binfeng Jia, and Shuangyang Wang. 2023. Out of the Box Thinking: Improving Customer Lifetime Value Modelling via Expert Routing and Game Whale Detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3206–3215.
- [55] Shiwei Zhao, Runze Wu, Jianrong Tao, Manhu Qu, Minghao Zhao, Changjie Fan, and Hongke Zhao. 2023. perCLTV: A general system for personalized customer lifetime value prediction in online games. *ACM Transactions on Information Systems* 41, 1 (2023), 1–29.