A Difficulty-Aware Framework for Churn Prediction and Intervention in Games

Jiayu Li[1], Hongyu Lu[1], Chenyang Wang[1], Weizhi Ma[1], Min Zhang[1]∗, Xiangyu Zhao[2], Wei Qi[2], Yiqun Liu[1], and Shaoping Ma[1]
[1]Department of Computer Science and Technology, Institute for Artificial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China
[2]Beijing Microfun Co.Ltd
{jy-li20,luhy16,wangcy18}@mails.tsinghua.edu.cn,mawz12@hotmail.com,{z-m,yiqunliu,msp}@tsinghua.edu.cn
{xiangyu.zhao,victor.qi}@microfun.com

ABSTRACT

User's leaving from the system without further return, called user churn, is a severe negative signal in online games. Therefore, churn prediction and intervention are of great value for improving players' experiences and system performance. However, the problem has not been well-studied in the game scenario. Especially, some crucial factors, such as game difficulty, have not been considered for large-scale churn analysis. In this paper, a novel Difficulty-Aware Framework (DAF) for churn prediction and intervention is proposed. Firstly, a Difficulty Flow for each user is proposed, which is utilized to derive users' Personalized Perceived Difficulty during the game process. Then, a survival analysis model *D-Cox-Time* is designed to model the Dynamic Influence of Perceived Difficulty on player churn intention. Finally, the *Personalized Perceived Difficulty* (PPD) and *Dynamic Difficulty Influence* (DDI) are incorporated to churn prediction and intervention. The proposed DAF framework has been specified in a real-world puzzle game as an example for churn prediction and intervention. Extensive offline experiments show significant improvements in churn prediction by introducing difficulty-related features. Besides, we conduct an online intervention system to adjust difficulty dynamically in the online game. A/B test results verify that the proposed intervention system enhances user retention and engagement significantly. To the best of our knowledge, it is the first framework in games that illustrates an in-depth understanding and leveraging dynamic and personalized perceived difficulty during game playing, which is easy to be integrated with various churn prediction and intervention models.

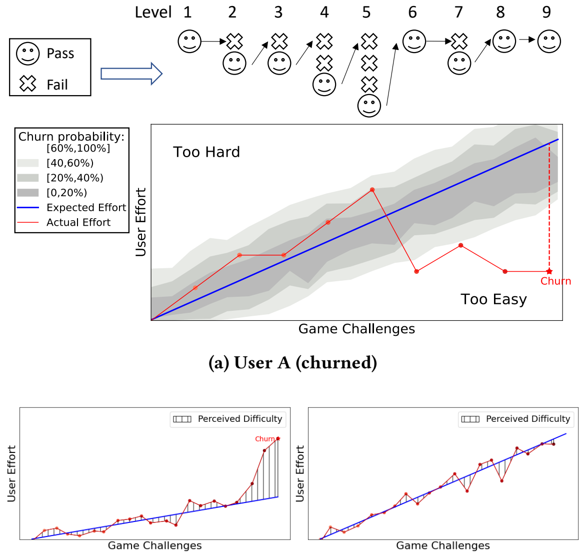
CCS CONCEPTS

• **Information systems** → **Data management systems**; **Personalization**; *Web log analysis*; • **Computing methodologies** → **Neural networks**.

KEYWORDS

Game difficulty; Churn prediction; Churn Intervention

ACM Reference Format:
Jiayuli, Hongyu Lu, Chenyang Wang, Weizhi Ma, Min Zhang, Xiangyu Zhao, Wei Qi, Yiqun Liu, and Shaoping Ma. 2021. A Difficulty-Aware Framework for Churn Prediction and Intervention in Games . In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3447548.3467277



(a) User A (churned)

(b) User B (churned)

(c) User C (retained)

Figure 1: An illustration of perceived difficulty and its dynamic influence on churning. The *User Effort* indicates time/experience that the user pay for the game. And the *Game Challenges* denotes global difficulty of the game.

1 INTRODUCTION

User Churn is a severe negative user feedback in online games, which means the user leaves this system and will not return for a long time. Hence effective user churn prediction and intervention are crucial in game scenarios.

Previous churn prediction studies in game scenario either apply existing methods with churn-related features in game [3, 4, 23], or analyze players' overall behaviors [17, 32]. However, they pay little attention to understanding and applying game-specific factors, such as game difficulty, for churn prediction. For churn intervention, adjustment difficulty is used intuitively. However, dynamic difficulty adjustment is usually conducted depending on explicit user feedback via questionnaires or manually constructed rules [6, 20, 31], which is impractical for real-world applications.

Therefore, we aim to analyze the difficulty factor in game thoroughly for predicting and preventing user churn with large-scale

∗Min Zhang is the corresponding author.

logs in a general framework. To model player's perception of difficulty, previous work has measured it by the efforts (e.g. time/trials in the game) by users [2, 6, 13]. However, as the expectation of effort is personalized, the perceived difficulty of each user can be distinct. For instance, spending the same time, casual players will lose patience quickly, while serious players with a clear goal to achieve will endure it for leveling up.

Further, the influence of difficulty on user churn intention is dynamic, varying with different phases in playing [21]. Intuitively, receiving the same perceived difficulty will lead to different impact on user churn. For example, beginners are more sensitive to difficulty, easily deterred by failures in games. As players get familiar with the game, their tolerance for difficulty may promote.

Figure 1 illustrates the modeling of *Personalized Perceived Difficulty* (PPD) and its *Dynamic Difficulty Influence* (DDI). From the behavior logs of player, user efforts and corresponding game challenges are extracted and fit a *Personalized Difficulty Flow* to obtain expected efforts, as shown by the blue lines in the figure. Then, PPD is obtained by the gap between actual and expected efforts (Figure 1(b,c)). Users may churn because the game is too easy (User A) or too hard (User B), and the shadow in Figure 1(a) indicates the churn probability (i.e. DDI). Generally, churn probability is high when PPD is extreme, and the influence is changing with time.

Based on the above motivation, we propose that difficulty, especially the perceived difficulty and its dynamic influences for each user, should be considered for churn prediction and intervention. Therefore, we design a novel Difficulty-Aware Framework (**DAF**). The basic idea of the framework is to propose a general method to generate difficulty-related features. Then, the features will contribute to both churn prediction and intervention tasks in game.

The proposed framework mainly consists of three modules: *Personalized Perceived Difficulty* (PPD) modeling, *Dynamic Difficulty Influence* (DDI) modeling, and *Churn prediction and intervention*. Firstly, the PPD is generated from the gap between actual efforts and expected efforts from user interaction history, where *Flow theory* [27] is used for modeling personalized differences in difficulty perceiving. Then, the *Dynamic Difficulty Influence* of PPD on user churn at different phases is modeled. Inspired by survival analysis, we propose a time-dependent survival model, *D-Cox-Time*, which incorporate PPD in modeling probability of churn through time, and reflect the DDI in our framework. At last, PPD and DDI-related features are applied to *churn prediction and intervention* tasks.

A specification of the proposed framework for a real-world game is conducted, and experimental results on both churn prediction and intervention tasks achieved good performance. Offline experiments of churn prediction show that difficulty-related information improves prediction performance significantly. The online A/B test results confirm that the intervention on difficulty optimizes user retention and engagement significantly.

To sum up, our main contributions are as following:

- We are the first to propose a Difficult-Aware Framework for churn prediction and interventions in games, in which the *Personal Perceived Difficulty* and its *Dynamic Difficulty Influence* are applied to generate features for both churn prediction and intervention tasks. The framework is capable for working for different online games.
- In the proposed framework, *personalized difficulty flow* and a time-dependent survival analysis model *D-Cox-Time* are designed to detect users' personalized difficulty and model its dynamic influence for churn.
- A specialization of the framework for a real-world game is conducted, which performs well on both churn prediction task and online churn intervention task. Open source of offline dataset is in https://github.com/THUIR/DAF-for-churn. To the best of our knowledge, this is the first large-scale public available dataset involving detailed interaction logs in the game scenarios.

2 RELATED WORK

2.1 Difficulty in Games

Difficulty is one of the essential concepts in game [25]. Many researchers have studied the definition and detection of it. Juul [13] conducted offline and online user studies in video game to detect the complex meaning of difficulty. Constant and Levieux [6] studied the relation between objective and subjective evaluation of difficulty. Recently, neural networks were also used to detect global difficulty in game [22]. Based on the detection of difficulty, various user studies in psychology and computer science scenarios were conducted to inspect its influence on user motivation and engagement [17, 21]. These works clarify the definition of difficulty in game, and verify strong relation between difficulty and user experience.

Another view for detecting difficulty in game comes from the flow theory. *Flow* is a commonly used concept in psychology, which describes the mental state of full immersion and focus while doing a task [8]. In 2005, Dweck and Elliot [9] postulated the *Flow Theory* that one must have a delicate balance between the time/skill required (i.e., challenges or difficulty) in the task and their own skills in order to achieve the *Flow* state. In the game scenario, some small-scale user studies have indicated the *Flow* exists in the video games [7, 27], and out-of-flow interactions will influence player's confidence [6].

However, these works usually ignore the personalized experiences of difficulty in games. Futhermore, none of them tried to model the *Flow* from user interactions explicitly. Differently, our framework focuses on modeling personalized perceived difficulty from historical interactions of each user.

2.2 Churn Prediction in Games

Prediction of churn is a crucial task in game scenario, where users' history log data is modeled with various classification algorithms. Some works combined existing models with churn-related features in game. Traditional classifiers with some game-specific features were applied to predict churn in games [3, 23]. Bonometti et al. [4] employed deep neural networks and proposed a Bifurcating Model Framework to model early user-game interactions. Liu et al. [16] proposed a deep semi-supervised model to analyze micro-level churn prediction and macro-level churn rankings in large gaming platforms. Others conducted valuable analysis on user behaviors in game to explain the predictions. Lomas et al. [17] inspected the relationship between challenge and user engagement. Yang et al. [32] conducted user clustering by early behaviors and made predictions with user type attention network.

Table 1: Notations of primary concepts used in this paper.

Notation	Explanation
$\vec{e_i}$	Actual efforts sequence for user u_i. [1]
$\vec{\hat{e}_i}$	Expected efforts sequence for user u_i.
$\vec{c_i}$	Game challenges sequence for user u_i.
$\boldsymbol{X_i}$	Other basic covariates for user u_i.
$\vec{D_i}$	*Personalized Perceived Difficulty* sequence for user u_i.
$\vec{\beta}(t)$	*Dynamic Difficulty Influence* vector at time t.
h(t)	Churn risk by survival analysis at time t.

[1] The jth dimension of the vector is represented with a second subscript, i.e. $\vec{e}_{i,j}$.

Churn prediction is also widely studied in various scenarios, such as social applications [32], telecommunication [1], and financial services [30]. Usually, popular machine learning methods are integrated with scenario-specific elements for churn prediction in these works.

Instead of proposing one specific churn prediction model in game, we focus on modeling the important factors, personalized perceived difficulty and its dynamic influence, and propose a framework that can be incorporated into various models.

2.3 Churn Intervention in Games

Besides churn prediction, churn intervention is a further task in application, which aims to optimize user retention by adjusting features in the games. Considering the operational feasibility, previous works usually conducted intervention by dynamic difficulty adjustment (DDA). In DDA, game designers aim to keep players engaged continuously by balancing an accurate level of difficulty. Early works employed probabilistic models in DDA and tried to gain local optimum for user engagement [5, 31]. Later, methods from machine learning are applied to DDA. For instance, reinforcement learning is used to develop adaptive AI in games [26, 28]. Recently, Pfau et al. [20] present a Deep Player Behavior Modeling strategy to adjust the difficulty with a more complex deep model.

However, these adjustment methods usually need ground truth for difficulty, or depend on manually designed assumptions on specific features about game, which is impractical for large-scale online game applications. In our work, the *Dynamic Difficulty Influence* is used for adjusting difficulty in real time for online scenario.

3 DIFFICULTY-AWARE FRAMEWORK FOR CHURN

3.1 Task Definition and Notations

We give the definitions of the two tasks and notations firstly.

Churn Prediction Task. Given user u_i and interaction behavior history $\vec{B_i} = b_{i,1}, b_{i,2}, ..., b_{i,T_o}$ in the observation window of length T_o, the churn prediction task is to predict the probability $P(churn|\vec{B_i})$ that u_i will churn, i.e. have no interactions in the detection window period T_d. The window size T_o and T_d are pre-defined due to specific scenario, which can vary from one day to a month.

Churn Intervention Task. Given behavior status b_i for user u_i and an action space A for all possible adjustment in system, the churn intervention aims to minimize the churn probability by conducting the proper action from A: $a = \arg\min_{a \in A} P(churn|b_i, a)$.

Notations. The main concepts used in this paper are displayed in the Table 1. Detailed derivation and explanation of these concepts are introduced in the following subsections.

3.2 Framework Overview

We design the **D**ifficulty-**A**ware **F**ramework for churn prediction and intervention (**DAF**), which aims at predicting and intervening player churn in games by modeling personalized difficulty and its influence on churn.

The overall structure of **DAF** is shown in Figure 2. The framework consists of three modules: *Personalized Perceived Difficulty* (PPD) modeling module, *Dynamic Difficulty Influence* (DDI) modeling module, and *Churn prediction and intervention* module. Here we give an overview of them, and the details of the three modules will be introduced later.

In PPD modeling module, given historical interaction logs of a player, game challenges and user efforts are extracted for modeling a *Personalized Difficulty Flow*, which outputs the PPD sequence $\vec{D}$. Then, in DDI modeling module, $\vec{D}$ are embedded and integrated with basic churn-related features along time to compute dynamic influence on churn. The DDI model is based on *D-Cox-Time*, a difficulty based variant of *Cox-Time* [15]. It includes parameter matrixes $A \in \mathbb{R}^{PxT}$ and $B \in \mathbb{R}^{MxT}$ to describe the global dynamic influence from basic features (of size P) and perceived difficulty (of size M) on churn hazard from time 1 to T, respectively. Dynamic influences, along with basic hazard $h_0(t)$, are updated with inputs and churn labels of every user. DDI are derived from B. Finally, PPD and corresponding DDI are incorporated in churn prediction classifiers and churn intervention system.

With pre-defined clarifications for each module (described in each section) based on game settings, our framework can be applied to various online games. As an example, a puzzle game specification of **DAF** is described in Section 4.

3.3 Personalized Perceived Difficulty Modeling

At the first step, we try to model users' PPD with the gap between their actual and expected efforts from historical interactions, as shown in the top-middle part of Figure 2.

With interaction logs of a user as input, user efforts and game challenges are extracted based on predefined specifications. Then the linear relation between user efforts and the game challenges is modeled with *Personalized Difficulty Flow*, which is used to generate personalized expected efforts. The gaps between users' actual effort and expected efforts form the output PPD.

To formalize, for user u_i, given the historical sequence of efforts $\vec{e_i}$ on all levels she has played, and the corresponding game challenges $\vec{c_i}$, linear regression is applied to model the *Flow*

$$\vec{e_i} = a_i * \vec{d_i} + b_i \tag{1}$$

Where a_i and b_i are the slope and intercept to describe the *Personalized Difficulty Flow* for u_i. Then, at any level $d_{i,j}$, we can obtain the personalized expected effort from the *Personalized Difficulty Flow*, $\hat{e}_{i,j} = a_i * c_{i,j} + b_i$. PPD $D_{i,j}$ is denoted by the difference between user's actual effort $e_{i,j}$ and expected effort $\hat{e}_{i,j}$:

$$PD_{i,j} = e_{i,j} - \hat{e}_{i,j} = e_{i,j} - (a_i * c_{i,j} + b_i) \tag{2}$$

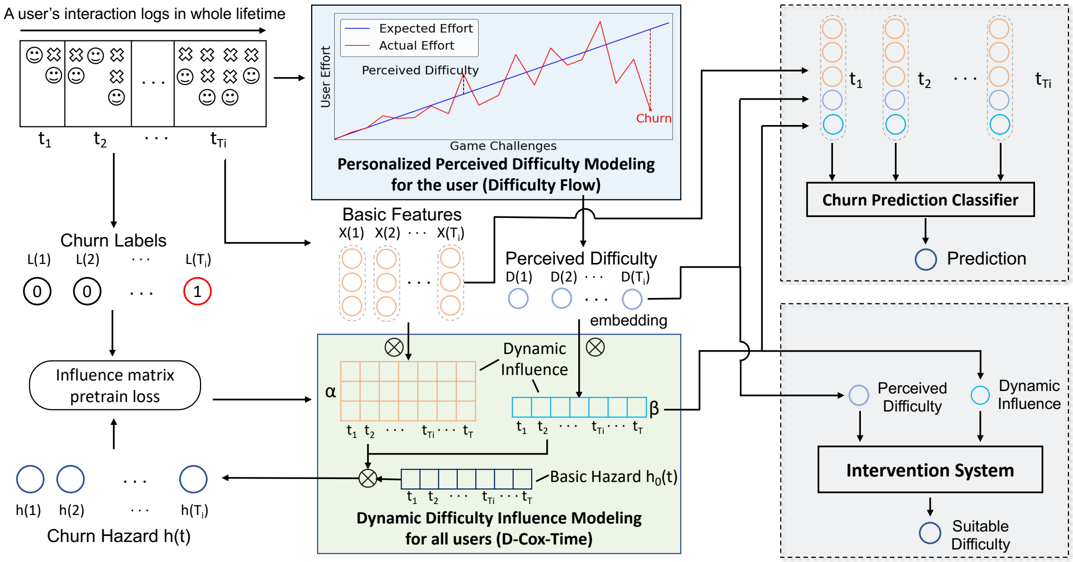


Figure 2: The Difficulty-Aware Framework (DAF) for churn prediction and intervention. ***Personalized Perceived Difficulty*** **is modeled with** ***Personalized Difficulty Flow*****, and** ***Dynamic Difficulty Influence*** **is obtained from** ***D-Cox-Time*** **model. Then the information about difficulty is incorporated in churn prediction and intervention tasks.**

Specification needs to be defined for efforts e and game challenges c in applications. The user effort e reflects personalized experience, such as time, money, or trials at the level. The game challenge c indicates global difficulty, such as level, average effort of all users, or manually-designed difficulty scales.

3.4 Dynamic Difficulty Influence Modeling

With a clear definition of PPD, we further take into consideration the *Dynamic Difficulty Influence* on players' churn. To perceive an explainable result, we propose a time-dependent difficulty-aware survival analysis model, *D-Cox-Time*, to describe the DDI, as shown in the bottom-middle in Figure 2.

Survival analysis is a common methodology for time-to-event prediction, modeling the risk of the event dynamically [14]. The goal for survival analysis is to model the probability of event happening as a function of time $P(T^* \le t)$, where T^* is the event happening time. Generally, the event *hazard rate* $h(t)$ is commonly used to specify models, which is defined by the derivative of probability $P(T^* \le t)$. For an in-depth overview of survival analysis, please refer to the book by Klein and Moeschberger [14].

For modeling the dynamic influence in **DAF**, we conduct a difficulty-aware time-dependent survival model *D-Cox-Time*, following the model *Cox-Time* [15]. Grain for time points can vary with applications, such as interaction-level, session-level, and even day-level. To describe delicate influence of PPD on churn at each time point, we conduct discretization and one-hot embedding for PD_{i,l_t} into pre-defined M dimensions, where l_t is the level that user u_i plays at time t.

Given the length of time sequence T, embedding for $\vec{D}_i \in \mathbb{R}^{M*T}$, basic covariates $\boldsymbol{X_i} \in \mathbb{R}^{P*T}$ (predefined information related to churn), the *D-Cox-Time* estimates the dynamic churn hazard for user u_i at time t ($t \le T$) as $h(t|D_{i,l_t}, \boldsymbol{X_i})$:

$$\begin{aligned} h(t|\boldsymbol{X_i}, D_{i,l_t}) = h_0(t) * exp\{ &\sum_{j=1}^{M} \mathbb{I}_{[\frac{j-1}{M},\frac{j}{M}]}(D_{i,l_t}) * \beta_j(t) \\ &+ \sum_{j=1}^{P} X_{i,j}(t)\alpha_j(t)\} \\ &:= h_0(t) * exp\{g(D_{i,l_t}, \vec{X}_i(t), \vec{\beta}(t), \vec{\alpha}(t)\} \end{aligned} \quad (3)$$

Where $\vec{h}_0 \in \mathbb{R}^T$, $\alpha \in \mathbb{R}^{P*T}$ and $\beta \in \mathbb{R}^{M*T}$ are parameters to learn. $h_0(t)$ is the global baseline hazard at time t, indicating the overall churn intention. $\beta_j(t)$ and $\alpha_j(t)$ are time varying influence on churn risk at time t, from each of the difficulty D_{i,l_t} and basic features $X_{i,j}$, separately. And $\mathbb{I}$ is the indicator function, illustrating the process of one-hot embedding.

At last, the DDI is indicated by the parameters $\beta_j(t)$, where larger value of $\beta_j(t)$ illustrates higher risk for churn, thus larger influence. A mapping function f is established between $\vec{D}$ and β:

$$f(D_{i,l_t}, t) = \beta_k(t) \quad (4)$$

Where k is the high bit in one-hot embedding of D_{i,l_t}.

For training *D-Cox-Time*, the loss function is defined

$$\begin{aligned} loss = \frac{1}{n} \sum_i log(\sum_{j \in \tilde{R}_i} exp[g(\vec{X}_j(t), D_{j,l_{T_i}}, \vec{\beta}(t), \vec{\alpha}(t)) \\ - g(D_{i,l_{T_i}}, \vec{X}_i(t), \vec{\beta}(t), \vec{\alpha}(t))]) \end{aligned} \quad (5)$$

Where T_i denotes the churn time for user u_i, and $\tilde{R}_i$ is a sampled subset of users *at risk* at time T_i (i.e. users not churned before time T_i). It has been proved that loss on $\tilde{R}_i$ is a good approximation of loss on all users at risk, and the loss on subset reduces the time complexity of the time-dependent model significantly[15].

To carry out the DDI modeling, settings that should be specification includes: grain for time points; maximum length T; basic churn-related covariates $\boldsymbol{X_i}$; one-hot embedding dimension M.

3.5 Churn Prediction and Intervention

In the end, we incorporate our in-depth modeling of PPD and DDI in churn prediction and churn intervention tasks, illustrated in the right part of Figure 2. As a framework, our methods can be used in various models for churn prediction and intervention.

Churn Prediction. Following definitions in Section 3.1, we formalize the churn prediction that considers difficulty information:

Given behavior history $\vec{B_i}$ for user u_i, the basic information $\boldsymbol{X_i}$, PPD $\vec{D_i}$, and DDI $\vec{\beta_i}$ are generated as features. Specifically, $\vec{\beta_i}$ indicates the corresponding *Dynamic Difficulty Influence* to $\vec{PD_i}$ calculated from mapping function f (equation 4). Then we predict the churn probability with a classifier M for different combinations of features: $P_M(churn|\boldsymbol{X_i})$, $P_M(churn|\boldsymbol{X_i}, \vec{PD_i})$, and $P_M(churn|\boldsymbol{X_i}, \vec{PD_i}, \vec{\beta_i})$.

The accuracy of prediction with different feature combinations can be calculated with various existing classifiers to verify the reliability of difficulty-aware modeling in **DAF**.

Churn Intervention. For churn intervention, we aim at providing proper difficulty to optimize user retention and engagement. Since the intervention is personalized and conducted in real time, *Personalized Difficulty Flow* is fit with early behaviors of each user (eq 1). And the DDI $\boldsymbol{\beta}$ is also pre-trained with *D-Cox-Time*.

Then, we perform the churn intervention system in a greedy way, i.e., minimizing churn hazard $h(t)$ at each time point t.

For user u_i, given an adjustment space for user effort at time t $e_{i,t}$: $E = \{e_1, e_2, ...e_n\}$, the best effort is optimized with

$$\begin{aligned} e_{i,t} &= \arg\min_{e_k \in E} h(t|D_{i,l_t}, \boldsymbol{X_i}) = \arg\min_{e_k \in E} g(D_{i,l_t}, \vec{X_i}(t), \beta(t), \alpha(t)) \\ &= \arg\min_{e_k \in E} f(D_{i,l_t}, t) = \arg\min_{e_k \in E} f(e_k - (a_i \cdot c_i(t) + b_i), t) \end{aligned} \tag{6}$$

The adjustment space E is defined by game designers. For instance, it can be difficulty scales at the level, the number of help provided in the level, or success and failure at each trial.

The churn intervention system is flexible to expand to various online games, and it is well-adapted for real-time adjustment as the computational complexity is low.

4 SPECIFICATION IN A TILE-MATCHING PUZZLE GAME

As illustrated in Section 3, **DAF** is easy to incorporate in various games. In this section, we conduct a specification of **DAF** in a real-world tile-matching game. Definition of concepts in difficulty-aware modeling is meanly displayed in Section 4.2, and experimental settings for churn are proposed in Section 4.3.

4.1 Dataset Collection

We collected anonymous data from a real-world tile-matching puzzle mobile game. This game is split to more than a thousand levels, which must be completed in sequence. At each level, users are posed with a puzzle and a goal (e.g., achieving a minimum score in a fixed period). If the players meet the goal, they will unlock the next level. Otherwise, they will lose energy in game and have to try again.

5000 new players who registered from Feb 1, 2020 to Feb 8, 2020 were selected, and their behavior data in the next two months was collected. After filtering players with less than 3 days of login data, 4089 users remained. There are 4,517,349 interaction records in the behavior logs in total, 1104 interactions for each user on average. The highest level played by user varies from 5 levels to 988 levels, with around 224 levels on average. Since the challenges in this game are represented by different levels, we mix the statement of "challenge" and "level" in the following description.

After removing the sensitive information about user privacy, we provide the behavior logs and codes of our experiments, to support further researches on games and difficulty [1].

4.2 Specification of DAF

The grain for time points is set to day-level for *D-Cox-Time* and churn prediction, since we consider the daily user churn event in the following experiment. The period of detection window is set as $T_d = 7$ days and observation window $T_o = 30$ days. Under this setting, there are 2239 churned users and 1850 retained users.

4.2.1 *Specification on Personalized Perceived Difficulty.*

According to Section 3.3, we define the game challenges and user efforts to specify PPD with *Personalized Difficulty Flow* in the dataset. Since users usually will not re-play a level once they pass it, we use the retry times (i.e., number of playing until success) at each level to represent the difficulty, following previous works in difficulty analysis [31]. The game challenge at each level is indicated with average retry times of all users who have played the level. The higher the average retry times are, the more challenging the level should be. The user effort is represented by the user's own retry times, which reflects her effort paid at the level.

4.2.2 *Specification on Dynamic Difficulty Influence.*

As illustrated in Section 3.4, to specify *D-Cox-Time* on the dataset, we clarify definition of basic covariates $\boldsymbol{X_i}$ and embedding dimension M for D_i. Since we consider the day-level user churn event in our experiment, covariate features $\boldsymbol{X_i}$ for *D-Cox-Time* is extracted from each day of logs. To be specific, $P = 14$ basic features are collected for each day from user behavior logs (as shown in Table 2), most of which are not game-specific. Following previous works on games [19], we categorize these features into five groups, and the period for ending a *Session* is set as 30 minutes without interactions.

Average of PPD at all levels at day t is calculated for $D_{i,t}$. Embedding dimension M is set to 10.

Finally, the DDI is represented by $\beta_j(t)$, where $j \in [1, M]$ and $t \in [1, T]$ ($M = 10, T = 30$).

4.3 Experimental settings

Experimental settings for churn prediction and intervention are determined according to definitions in Section 3.5.

4.3.1 *Settings for Churn Prediction.*

Features, classifiers, and evaluation protocols are set for churn prediction.

[1] https://github.com/THUIR/DAF-for-churn

Table 2: Basic Covariates for survival analysis in *D-Cox-Time.*

Feature Type	Feature	Description
Playing intensity	Session num	Number of sessions in the day.
	Played num	Number of plays in the day/last session
	Last session play	
	Played levels	Number of levels played in the day/last session
	Last session level	
Player attention	Game time	Play time (seconds) in the day/last session
	Last session time	
	Session length	Average number of plays per session in the day
Player loyalty	Help num	Number of helps used in the day/last session
	Last session help	
	Purchase amount	Amount of purchase in the day
Context	Weekday	Day of week
	Last session end hour	End hour of the last session.
Player level	Player level	Highest level in the day

Table 3 summarizes the features for churn prediction. *Basic* feature categories are almost the same as basic covariates in $D-Cox-Time$ (Table 2), except that the features are pooled in the observation window for non-sequence prediction models. (Average over all days and sum of purchase num are considered in pooling), and the original sequence is used for sequential classifiers. The *Difficulty* category contains information from modeling of *Difficulty Flow*: the game challenges, user efforts, and PPD. And the DDI category considers information from influence matrix $\boldsymbol{\beta}$ of *D-Cox-Time.*

As our main goal is to verify the difficulty-related features, but not to propose new models for churn prediction, we follow the previous works[23, 24, 29], and use four types of traditional classification models for the churn prediction task:

1) **Basic model**: logistic regression (LR) and support vector machines (SVM).

2) **Deep model**: MultiLayer Perceptron (MLP) and DeepFM [11].

3) **Sequential model**: Long Short-Term Memory(LSTM) [10]. (Feature sequence of each day in user behavior logs is used as input.)

4) **Ensemble model**: random forest (RF) and Gradient Boosting Decision Tree (GBDT).

Five-fold cross-validation is conducted for evaluation. *D-Cox-Time* is pretrained only on training set. AUC, accuracy (ACC), and F1 value are used as evaluation metrics.

4.3.2 Online Intervention System Construction. For online churn intervention, we make some modifications on pretrain of PPD an DDI, and then adjustment space E is specified.

We focus on intervention for new users in the online experiment. The PPD for each new user is modeled with their interactions at beginning. Since intervention should be conducted at each interaction, *D-Cox-Time* is trained on interaction grain for DDI on online data of 20,000 new users, and the end of session is set as label.

With the pre-trained *Difficulty Flow* for PPD and *D-Cox-Time* for DDI, adjustment space E consists two states: $E = \{e_p, e_f\}$, e_p for passing the current level, and e_f for failing. The proper state at each interaction is chosen by optimization in Equation 6.

A/B test [12] is used for online performance evaluation. Experimental group and control group randomly selected 30,000 new users from the system, respectively who registered at the same day. During a period of 10 days, users in experimental group received the intervention based on awareness of difficulty by the proposed **DAF**, and users in control group have no adjustment. The settings for experimental group and control group are all the same.

Experimental results will be shown in Section 6.

5 ANALYSIS ON DIFFICULTY MODELING

In this section, we analyse the PPD and DDI from difficulty modeling of the offline dataset.

5.1 Analysis on Personalized Perceived Difficulty

According to Equation 1, the *Personalized Difficulty Flow*s for all users are fit from their historical interactions in the 30-day observation window under settings in Section 3.3. The average coefficient of determination $R^2 = 0.617$ (high goodness of fit[18]), indicating a good linear relation between game challenges and user efforts.

The distribution of personalized slopes a_i is shown in Figure 3. Users are grouped into three phases according to their highest level. The distribution, peak value, mean value (μ), and variance (σ^2) are shown. A smaller value of a indicates a preference for easier games.

For most users, the slopes a for *Personalized Difficulty Flow* are positive, and the distribution of a is clumping to median as players get experienced. It indicates there are positive proportional correlations between effort and game challenge for most users in all phases, centralizing slopes in the range of (0,1), which confirms the settings of *Flow*. Distributions in all three phases are single-peaked. As users get familiar with the game, the peak of distribution is approaching its mean value, and the variance of slopes is decreasing. At fresh-man phase, the peak value is small, and players' *Personalized Difficulty Flow*s are diverse. It illustrates generally players prefer easier games at the beginning, but players who prefer easy or hard games are both at scale. As players get experienced at the game, most users tend to have a balanced relation between challenges and efforts, and the distribution is approaching a Normal Distribution.

5.2 Analysis on Dynamic Difficulty Influence

We perform training of *D-Cox-Time* in Section 3.4 on the whole offline dataset, and analyze β from overall and time-various aspects.

First, the mean values and variances for DDI, i.e $\beta_j(t)$ are shown at $M = 10$ different bins of PPD in Figure 4. Larger value of β indicate greater probability to churn. It illustrates the tendency that the average churn risk is lower when the PPD is close to 0, i.e., the user's actual effort matches her expected effort. And the hazard is high when PPD is too small (too easy) or too large (too difficult).

Then we inspect how the DDI $\boldsymbol{\beta}$s change through time. To better understand the shift, the $M = 10$ PPD bins are grouped into 3 ranges, and the observation window is divided into three phases: less than 10 days, 10 days to 20 days, and more than 20 days.

The average value for $\boldsymbol{\beta}$ along PPD and time are shown in Table 4. It illustrates that the influences change through time in all ranges

Table 3: Features for churn prediction. For sequential models in Section 4.3, the features are extracted per day. For other models, the day-level features are averaged through the whole observation window.

Category	Feature Type	Feature	Description
Basic	Playing intensity	Session num, Played num, Played levels, Login Frequency	Same to Table 1.
	Player attention	Game time	Same to Table 1.
	Player loyalty	Help num, Purchase num	Same to Table 1.
		First purchase interval, Last purchase interval	Days from registration to the first/last purchase.
	Player level	Player level	Highest level ever played.
Difficulty	Game Challenges	Average game challenges, game challenges variance, Last-day game challenges	Mean and variance of game challenges per day / at the day before churn.
	User Efforts	Average effort, Effort variance, Last-day effort	Mean and variance of effort per day / at the day before churn.
	PPD	Average PPD, PPD variance, Last-day PPD	Mean and variance of PPD per day / at the day before churn.
DDI	Beta	Average Beta, Beta variance, Last-day Beta	Mean and variance of Beta coefficients per day / at the day before churn.

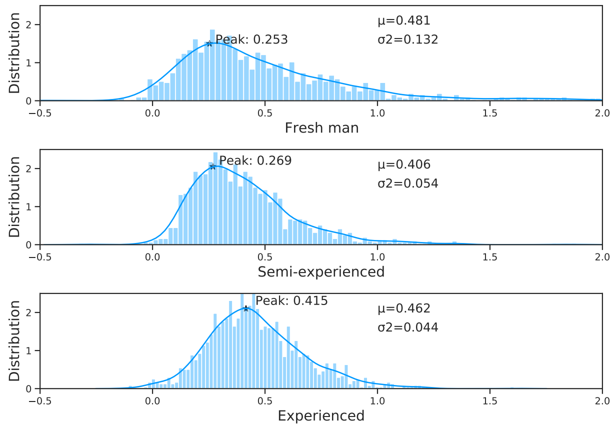


Figure 3: Distribution of slope a for users in three phases. The distribution of slopes centralizes in the range (0,1).

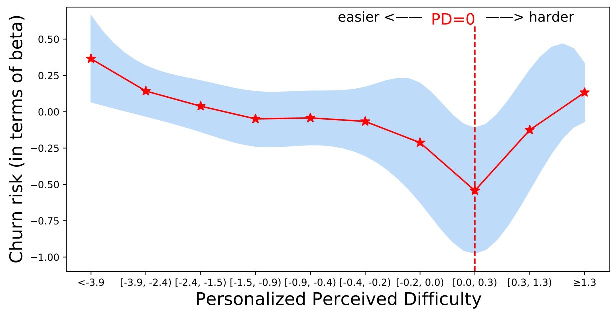


Figure 4: Average and variances of churn risk (in terms of β, eq 3) for each bins for *Personalized Perceived Difficulty*. The churn risk is high when PPD is too low or too high.

Table 4: The change of churn risk(in terms of β) through time, where higher value means higher risk to churn.

	PD < -0.4	-0.4≤PD<1.3	PD≥1.3
<10 days	0.053	0.081	0.160
10 days - 20 days	0.230	-0.065	0.149
≥20 days	0.183	-0.542	0.272

of *PD*s. From the specific values, it can be found that when actual effort matches the user's expected effort (i.e., *PD* near 0), the user is less likely to churn, especially at long phase. At the beginning, feeling easy or exact matching has little influence on churn, but too difficult games increase users' probability to churn. Later, feeling too easy or too difficult will both lead users to churn, while in the median phase, feeling easy is riskier to churn, but hard games are more harmful in the longer phase.

6 CHURN PREDICTION & INTERVENTION RESULTS

6.1 Churn Prediction Results

For churn prediction, hyper-parameters of all seven models in Section 4.3 are carefully tuned on the dataset, and the best hyper-parameters are recorded in our codes[1]. Results with the highest AUC value are reported in Table 5.

From the results, it can be concluded that adding features in *Difficulty* and *DDI* categories improve the performance of prediction models. The improvements of AUC are significant for both categories in almost all models except LR. Therefore, the results indicate that incorporating the proposed *Personalized Perceived Difficulty* and *Dynamic Difficulty Influence* can improve the prediction performance for user churn. Especially, in LSTM, adding features in *DDI* enhance the performance the most among all models, from 0.913 to 0.953 (+4.0%), which indicates that *DDI* is suitable for prediction of sequential model. This may because the DDI information comes

[1] https://github.com/THUIR/DAF-for-churn

Table 5: Overall performance of different models on different features. '+ DIF' means adding features in the *Difficulty* category. '+ DDI' indicates adding features in the DDI category. Paired t-test is conducted on the results. * and ** denote the statistical significance for $p < 0.05$ and $p < 0.01$ respectively, compared to the previous feature group in the same model. And the models are optimized with the best AUC.

Model	Features	AUC	ACC	F1
LR	Basic	0.868	0.781	0.800
	Basic + DIF	0.908^{**}	0.837^{**}	0.848^{**}
	Basic + DIF + DDI	0.914^{*}	0.842	0.851
SVM	Basic	0.834	0.834	0.846
	Basic + DIF	0.878^{**}	0.875^{**}	0.881^{**}
	Basic + DIF + DDI	0.889^{*}	0.885^{*}	0.891^{*}
MLP	Basic	0.920	0.843	0.855
	Basic + DIF	0.940^{**}	0.875^{**}	0.884^{**}
	Basic + DIF + DDI	0.951^{**}	0.888^{*}	0.897
DeepFM	Basic	0.909	0.835	0.848
	Basic + DIF	0.935^{**}	0.877^{**}	0.885^{**}
	Basic + DIF + DDI	0.947^{*}	0.881	0.893^{*}
LSTM	Basic	0.902	0.816	0.820
	Basic + DIF	0.913^{*}	0.834^{*}	0.846^{*}
	Basic + DIF + DDI	0.953^{**}	0.874^{**}	0.877^{**}
RF	Basic	0.913	0.835	0.845
	Basic + DIF	0.933^{**}	0.863^{**}	0.867^{*}
	Basic + DIF + DDI	0.955^{*}	0.886^{*}	0.892^{*}
GBDT	Basic	0.969	0.919	0.925
	Basic + DIF	0.976^{*}	0.932^{*}	0.934^{*}
	Basic + DIF + DDI	**0.980^{*}**	**0.940^{**}**	**0.945^{*}**

from the time-dependent model *D-Cox-Time*, thus better describe the sequential changes.

Among all the models, GBDT performs the best. Even with the *Basic* features, GBDT performs better than all feature categories in other models. This may because ensemble methods are effective in our churn prediction task. Meanwhile, it can be found that deep models are not so powerful for the problem since we perform on a small dataset with about 4000 users.

Moreover, it is observed that although more information is used for training in the sequential model LSTM, it still performs worse than ensemble models. On the one hand, the reason might be that behavior sequence contains much noise, and users may have different patterns for churn, which is hard to be learned by a single RNN. On the other hand, it demonstrates that the features we extracted for non-sequential models are effective.

6.2 Further Analysis on Churn Prediction

As the influence of difficulty on churn is dynamic (See Table 4) and user lifetime is an essential metric in application, we further explore how difficulty helps churn prediction in different phases.

Following the analysis in Section 5.2, the 30-day observation window is divided into three phases, churning before 10 days, in 10 days to 20 days, and after 20 days. At each phase, churn user is defined the same as Section 3.5, and all other users who have not churned yet are labeled as retain users. We conduct churn prediction experiments under the three new settings with the best-performance model, GBDT, and the model is optimized with AUC.

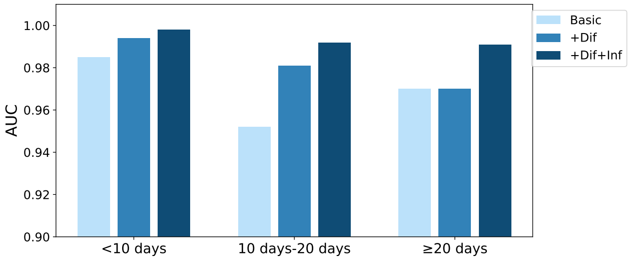


Figure 5: Churn prediction results (in terms of AUC) on three phases of users with GBDT model.

The churn prediction performances for different phases are shown in Figure 5. At any phase, adding difficulty-related features promotes performance. With all features, the GBDT gains AUC greater than 0.99 on all three datasets. Basic features perform well enough (AUC=0.985) when users churn before 10 days, but the performance declines in longer period. In 10-20 days, adding *Difficulty* features improves the AUC significantly. After 20 days, *Difficulty* features contribute less, but adding DDI features further improve the performance significantly.

Therefore, *Difficulty* and DDI help predict users' churn at any stage, and they are especially beneficial in longer periods.

6.3 Online Churn Intervention Results

In online experiment, the *Difficulty Flow* is fit with the first 30 levels for each user, and users with less than 30 levels are filtered. *D-Cox-Time* is trained with online data of 20,000 users before A/B test. In 10-day A/B test experiments, we collect 14,615 users in the experimental group and 14,687 users in the control group.

We consider four metrics commonly used in industry scenarios:

- *Next-day churn percentage*: Average percentage of users logging in at day *T* but without interactions at day *T+1* (T=1,2,...,9). Lower next-day churn percentage indicates better retention of users in short period.
- *Week-churn percentage*: Percentage of users who had interactions between day1 and day7, but have no interactions in the following day8 to day10. Lower week-churn percentage indicates better retention in long period.
- *Total playtime per user*: Average of total playtime for each user in ten days. Longer time indicates more engagement.
- *Average session length*: Average amount of interactions in the sessions. Longer session length denotes more engagement.

The A/B test results are shown in Table 6. The average next-day churn percentage decreases 10.9%, and the longer week-churn percentage decreases nearly 20%, indicating the intervention system does optimize retention of users in the game. Moreover, the engagement for users has significant improvement. On average, compared with control group, users in experimental group spend 10% more time on game, and have more interactions in each session.

Therefore, the difficulty intervention system for online application can improve user retention and engagement significantly, which further verify our Difficulty-Aware Framework is efficient.

Table 6: The online A/B test experiment results. To protect commercial privacy of the game company, we only present the relative gain between experimental group and control group. ↓ (↑) indicates lower (higher) means better performance. Independent t-test is conducted on the results, where ** denotes the statistical significance for $p < 0.01$.

Metric	Experimental Group vs. Control Group
Next-day churn percentage ↓	-10.9%**
Week-churn percentage ↓	-19.7%**
Total playtime per user ↑	+11.0%**
Average session length ↑	+4.6%**

7 CONCLUSIONS AND FUTURE WORK

In this work, we propose a Difficulty-Aware Framework for churn prediction and intervention, which provide an in-depth understanding of difficulty in online game scenarios. Based on user behavior history, we model the *Personalized Perceived Difficulty* (PPD) by difficulty flow, and its *Dynamic Difficulty Influence* (DDI) on churn is with a time-dependent survival analysis model *D-Cox-Time*. The analysis verifies that users follow a proportional relation between subjective effort and the objective game challenges, and too hard or too easy games will both lead to higher churn risk. Due to the flexibility of our framework, it can be applied to different games easily. In a specialization on a real-world online game, significant improvements are achieved with the extracted difficulty-related features. Difficulty-aware features enhanced models outperform the originals significantly in churn prediction. Besides, online A/B test shows the effectiveness of introducing difficulty features in churn intervention.

In the future, we plan to further investigate more complex distributions of *Personalized Perceived Difficulty* to understand the process of difficulty perception. Besides, the current DAF takes a two-steps strategy, i.e. difficulty learning and prediction/intervention. We will continue to investigate whether it is possible to conduct an end-to-end framework. Moreover, the difficulty modeling framework can be applied to other scenarios such as online education.

8 ACKNOWLEDGEMENTS

This work is supported by the National Key Research and Development Program of China (2018YFC0831900), Natural Science Foundation of China (Grant No. 62002191, 61672311, 61532011) and Tsinghua University Guoqiang Research Institute. This project is also funded by China Postdoctoral Science Foundation (2020M670339). Dr. Weizhi Ma is supported by Shuimu Tsinghua Scholar Program.

REFERENCES

[1] Jae-Hyeon Ahn, Sang-Pil Han, and Yung-Seop Lee. 2006. Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications policy* 30, 10-11 (2006), 552–568.

[2] Maria-Virginia Aponte, Guillaume Levieux, and Stephane Natkin. 2011. Measuring the level of difficulty in single player video games. *Entertainment Computing* 2, 4 (2011), 205–213.

[3] Paul Bertens, Anna Guitart, and África Periáñez. 2017. Games and big data: A scalable multi-dimensional churn prediction model. In *2017 IEEE conference on computational intelligence and games (CIG)*. IEEE, 33–36.

[4] Valerio Bonometti, Charles Ringer, Mark Hall, Alex R Wade, and Anders Drachen. 2019. Modelling Early User-Game Interactions for Joint Estimation of Survival Time and Churn Probability. In *2019 IEEE Conference on Games (CoG)*. IEEE, 1–8.

[5] Sara Bunian, Alessandro Canossa, Randy Colvin, and Magy Seif El-Nasr. 2017. Modeling individual differences in game behavior using HMM. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 13.

[6] Thomas Constant and Guillaume Levieux. 2019. Dynamic difficulty adjustment impact on players' confidence. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[7] Thomas Constant, Guillaume Levieux, Axel Buendia, and Stéphane Natkin. 2017. From objective to subjective difficulty evaluation in video games. In *IFIP Conference on Human-Computer Interaction*. Springer, 107–127.

[8] Mihaly Csikszentmihalyi and Mihaly Csikzentmihaly. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row New York.

[9] Carol S Dweck and Andrew J Elliot. 2005. *Handbook of competence and motivation*. Guilford Press New York.

[10] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with LSTM. (1999).

[11] 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).

[12] Katja Hofmann, Lihong Li, and Filip Radlinski. 2016. Online evaluation for information retrieval. *FnTIR* 10, 1 (2016), 1–117.

[13] Jesper Juul. 2009. Fear of failing? the many meanings of difficulty in video games. *The video game theory reader* 2, 01 (2009), 2009.

[14] John P Klein and Melvin L Moeschberger. 2006. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

[15] Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. 2019. Time-to-Event Prediction with Neural Networks and Cox Regression. *arXiv:1907.00825* (Sept. 2019). arXiv: 1907.00825.

[16] Xi Liu, Muhe Xie, Xidao Wen, Rui Chen, Yong Ge, Nick Duffield, and Na Wang. 2020. Micro-and macro-level churn analysis of large-scale mobile games. *Knowledge and Information Systems* 62, 4 (2020), 1465–1496.

[17] Derek Lomas, Kishan Patel, Jodi L Forlizzi, and Kenneth R Koedinger. 2013. Optimizing challenge in an educational game using large-scale design experiments. In *SIGCHI*. 89–98.

[18] Nico JD Nagelkerke et al. 1991. A note on a general definition of the coefficient of determination. *Biometrika* 78, 3 (1991), 691–692.

[19] Africa Perianez, Alain Saas, Anna Guitart, and Colin Magne. 2016. Churn Prediction in Mobile Social Games: Towards a Complete Assessment Using Survival Ensembles. In *DSAA*.

[20] Johannes Pfau, Jan David Smeddinck, and Rainer Malaka. 2020. Enemy within: Long-term motivation effects of deep player behavior models for dynamic difficulty adjustment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–10.

[21] Hua Qin, Pei-Luen Patrick Rau, and Gavriel Salvendy. 2010. Effects of different scenarios of game difficulty on player immersion. *Interacting with Computers* 22, 3 (2010), 230–239.

[22] Shaghayegh Roohi, Asko Relas, Jari Takatalo, Henri Heiskanen, and Perttu Hämäläinen. 2020. Predicting Game Difficulty and Churn Without Players. In *The Annual Symposium on Computer-Human Interaction in Play*. 585–593.

[23] Karsten Rothmeier, Nicolas Pflanzl, Joschka Hüllmann, and Mike Preuss. 2020. Prediction of Player Churn and Disengagement Based on User Activity Data of a Freemium Online Strategy Game. *IEEE Transactions on Games* (2020).

[24] Mehpara Saghir, Zeenat Bibi, Saba Bashir, and Farhan Hassan Khan. 2019. Churn prediction using neural network based individual and ensemble models. In *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. IEEE, 634–639.

[25] Katie Salen, Katie Salen Tekinbaş, and Eric Zimmerman. 2004. *Rules of play: Game design fundamentals*. MIT press.

[26] Yoones A Sekhavat. 2017. MPRL: Multiple-Periodic Reinforcement Learning for difficulty adjustment in rehabilitation games. In *2017 IEEE 5th international conference on serious games and applications for health (SeGAH)*. IEEE, 1–7.

[27] Penelope Sweetser and Peta Wyeth. 2005. GameFlow: a model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)* 3, 3 (2005), 3–3.

[28] Chin Hiong Tan, Kay Chen Tan, and Arthur Tay. 2011. Dynamic game difficulty scaling using adaptive behavior-based AI. *IEEE Transactions on Computational Intelligence and AI in Games* 3, 4 (2011), 289–301.

[29] Qiu-Feng Wang, Mirror Xu, and Amir Hussain. 2019. Large-scale ensemble model for customer churn prediction in search ads. *Cognitive Computation* 11, 2 (2019), 262–270.

[30] Artit Wangperawong, Cyrille Brun, Olav Laudy, and Rujikorn Pavasuthipaisit. 2016. Churn analysis using deep convolutional neural networks and autoencoders. *arXiv preprint arXiv:1604.05377* (2016).

[31] Su Xue, Meng Wu, John Kolen, Navid Aghdaie, and Kazi A Zaman. 2017. Dynamic difficulty adjustment for maximized engagement in digital games. In *Proceedings of the 26th International Conference on World Wide Web Companion*. 465–471.

[32] Carl Yang, Xiaolin Shi, Luo Jie, and Jiawei Han. 2018. I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application. In *SIGKDD*. 914–922.

REPRODUCIBILITY

To facilitate reproducibility of the results in this paper, we are sharing the data used at , and the codes at https://github.com/THUIR/DAF-for-churn. The open data contains raw user interactions and payment logs after removing the sensitive information about user privacy. The codes includes pre-processing of data, five-fold split results , modeling of difficulty, and churn prediction experiments. Hyper-parameters for all methods in churn prediction are included in the README.md for the codes.