Improved Retention Analysis in Freemium Role-Playing Games by Jointly Modeling Players' Motivation, Progression and Churn

Bikram Karmakar[1], Peng Liu[2], Gourab Mukherjee[3], Hai Che[4], Shantanu Dutta[5]

[1] Department of Statistics, University of Florida

[2] Department of Marketing, Santa Clara University

[3] Department of Data Sciences & Operations, University of Southern California

[4] Department of Marketing, UC Riverside

[5] Department of Marketing, University of Southern California

June 7, 2021

Abstract

We consider user retention analytics for online freemium role-playing games (RPGs). RPGs constitute a very popular genre of computer-based games that, along with a player's gaming actions, focus on the development of the player's in-game virtual character through a persistent exploration of the gaming environment. Most RPGs follow the freemium business model in which the gamers can play for free but they are charged for premium add-on amenities. As with other freemium products, RPGs suffer from the curse of high dropout rates. This makes retention analysis extremely important for successful operation and survival of their gaming portals. Here, we develop a disciplined statistical framework for retention analysis by modeling multiple in-game player characteristics along with the dropout probabilities. We capture players' motivations through engagement times, collaboration and achievement score at each level of the game, and jointly model them using a generalized linear mixed model (glmm) framework that further includes a time-to-event variable corresponding to churn. We capture the inter-dependencies in a player's level-wise engagement, collaboration, achievement with dropout through a shared parameter model. We illustrate interesting changes in player behaviors as the gaming level progresses. The parameters in our joint model were estimated by a Hamiltonian Monte Carlo algorithm which incorporated a divide-and-recombine approach for increased scalability in glmm estimation that was needed to accommodate our large longitudinal gaming data-set. By incorporating the level-wise changes in a player's motivations and using them for dropout rate prediction, our method greatly improves on state-of-the-art retention models. Based on data from a popular action based RPG, we demonstrate the competitive optimality of our proposed joint modeling approach by exhibiting its improved predictive performance over competitors. In particular,

The research was supported by NSF DMS-1811866 and NSF DMS-2015250 grants.

[3]Corresponding author: gmukherj@marshall.usc.edu

we outperform aggregate statistics based methods that ignore level-wise progressions as well as progression tracking non-joint model such as the Cox proportional hazards model. We also display improved predictions of popular marketing retention statistics and discuss how they can be used in managerial decision making.

Keywords: joint modeling, shared parameter model, freemium games, dropout, retention analysis, player motivation, consumer engagement, social contagion, role-playing games, divide and recombine, scalable glmm.

1 Introduction

Role-playing games (RPGs) constitute a very popular genre of computer-based games where the player controls the actions of a character in a persistent virtual world (Badrinarayanan et al., 2015, Chris, 1982, Pierre-Louis, 2019). He or she can assume a role of a fantasy character and interact with other players in the virtual game world (Bowman, 2010). RPGs have a devoted consumer base which is rapidly growing (Hill, 2019). As such, Statista (2018) predicts the revenue of RPG industry to reach 43 billion USD in 2021. A vibrant consumer base and huge revenue potential has attracted several game producers to this market segment. This has led to increased competition and the current gaming market is extremely crowded. Monetization policies associated with these digital products is rapidly revolutionizing the marketing and advertisement space in information systems (Appel et al., 2019, Liu et al., 2014). Another challenge faced by marketing managers is that most of the revenues from these games are generated through the *freemium business model* (Evans, 2016, Kumar, 2014, Niculescu and Wu, 2011) where consumers play the game for free and revenue is accumulated through advertisement or purchase of premium add-on components to the game. Developing a disciplined analytics framework for monetization and sustainability of these games is critical for these platforms.

In these freemium games there is no subscription, so dropout is a very important challenge that managers face (Yang and Peterson, 2004). Compared to other entertainment sectors, the dropout (churn) rate in freemium games is very high – typically ranging from 20% to 50% per month (Castro and Tsuzuki, 2015). Also, acquiring new players is costly and can significantly exceed the cost of retaining existing players (Fields, 2014). Thus, retention analysis, which involves predicting and understanding churn rates, is an important problem for revenue planning and budgeting in freemium RPG games (Kawale et al., 2009, Periáñez et al., 2016). Retention analysis is a widely used technique in a wide range of industries such as telecom (Mozer et al., 2000), retail business (Xie et al., 2009), banking (Coussement and Van den Poel, 2008), insurance (Morik and Köpcke, 2004) and credit card (Nie et al., 2009). However, dropout analysis in these new age of freemium products is intrinsically different from retention analysis in the aforementioned traditional business sectors (Borbora et al., 2011).

In freemium RPGs it is important to understand if there is a consistent rate of churn or if there are instances or game levels where there are higher dropouts. A player's propensity to continue playing the game is governed by his/her in-game experiences and achievements. These outcomes need to be factored in the dropout analysis. Here, we capture players' motivations through level-wise engagement times, collaboration and achievement score. We develop a disciplined statistical methodology for estimating the hazard rate of churn by tracking a player's behavior through the following three level-wise gaming attributes:

i. We measure a player's engagement E_l in the game at level l by the playing time needed by him/her to cover level l of the game. Estimating E_l for future players is important as high variations in these engagement times across different levels suggest non-uniformity in the game and can be potentially linked to peculiarities in the drop-out rate (Huang et al., 2019). These level-wise progression statistics E_l reflect the time-span in which a player is directly engaged in the game which provides the portal revenue earning opportunities such as banner ads.

ii. We measure a player's collaboration activity by the binary variable C_l which is positive if he/she has collaborated with other players and played at least one team game in level l. A distinguishing feature of RPGs is that they not only aspire for a player's skill development but also aim to provide an enhanced gaming experience by involving sharing, collaboration and team building in game-play. It has been witnessed that increased in-game social interactions and collaboration immensely help in adoption of the game and retaining players (Park et al., 2018, Wei et al., 2019, Zhang et al., 2017).

iii. We measure a player's achievement A_l at level l by the points he/she got by completing the level. In online RPGs, the level completion points vary across players depending on his/her exploits and achievement in the level. Rather than solely focusing on skill development, RPGs promote a never ending quest for exploration of the virtual world for the growth of the player's virtual character (Clements, 2012). A player gets satisfaction by fulfilling the minimal mission needs for level progression and also through additional quests that help to grow his/her in-game character. Players who get immersed in the game play more missions (and often more dangerous or more challenging ones) and quests than the average player. This is reflected in their above average level completion points (achievement score). It is easier to retain enthusiastic players (Castro and Tsuzuki, 2015) and so, it is important to understand the impact of these level-progression achievement scores in dropout rates.

We develop a joint modeling framework for simultaneous prediction of the aforementioned attributes as a player progresses through different levels of the game. Most popular retention metrics analysis including player's lifetime gaming characteristics (Gupta et al., 2004) such as the lifetime engagement (cumulative playing time across levels) or collaborations, can be easily computed as functions of these level-wise predictions.

The paper is organized as follows. In the following section, we briefly describe our joint modeling approach and its advantages over contemporary marketing retention methods. In Section 3, we describe the data set for our case example. In Section 4, we present a joint model for predicting level-wise playing characteristics (JMLPC) and describe the methodology for estimating the parameters in JMLPC. The implications from the estimated model, prediction results on the validation data set as well as a retention marketing application are provided in Section 5. In Section 6, we discuss several variants and extensions of our joint model approach and thereafter close with a discussion in Section 7.

2 Statistical Joint Modeling and Improved Retention Analysis

Engagement, collaboration and achievement are three pivotal disparate in-game measures that regulate a player's motivation for playing online RPG games (Borbora et al., 2011). In a statistical joint modeling framework (Rizopoulos and Lesaffre, 2014, Rizopoulos et al., 2010) we consider simultaneous estimation of the three longitudinal player motivation attributes via hierarchical general linear mixed models (Banerjee et al., 2014, Jiang, 2007). We track the joint progression of these player motivational characteristics and use their propagation as well as level invariant player attributes to model dropout as a time-to-event variable in the joint modeling framework (Rizopoulos, 2012, Rizopoulos et al., 2009). Figure 1 provides a schematic illustration of our analysis framework.

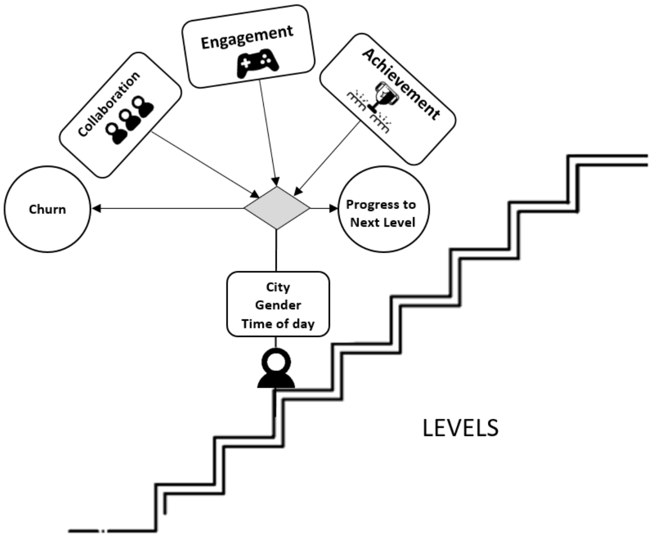


Figure 1: Schematic diagram of joint modeling based retention analysis framework.

Based on our estimated joint model JMLPC, we make predictions of the behaviors of future players. These predictions outperform predictions from non-joint model (see section 5.4) implying that modeling the co-dependencies between a player's motivations and churn rate is beneficial. Retention marketing is a well-established research area (East et al., 2006, Rosenberg and Czepiel, 1984). Though there are plenty of methods in the existing literature on churn prediction (see Borbora et al., 2011, Jerath et al., 2011, Periáñez et al., 2016 and the references therein), they do not model the co-dependencies between a player's in-game motivations. Here, by incorporating those dependencies via a joint modeling approach we provide improved predictions on future player behaviors. In sections 5.2–5.4, we show that better prediction on player behaviour also help in formulating better retention marketing policies. Using the estimated JMLPC model we can provide life-time player values by aggregating the player's expected playing time across levels. Marketers can use these estimates, particularly those pertaining to lifetime game-play volume, for efficient budgeting of advertisements in the gaming portal. We show that the predicted lifetime player values from our joint modeling set-up improve inference compared to predictions based on the marginal distributions (see section 5.4).

Another novel aspect of our study is that, unlike most recent approaches for analyzing player behavior in online games (Banerjee et al., 2019, Huang et al., 2019, Liu et al., 2020, Park et al., 2018), we consider modeling player behaviour aggregated for each level of the game and not at the daily or weekly resolutions. While daily active user (DAU) based key performance indicators (KPIs) show that higher levels increase a player's daily or weekly gaming involvements, they do not directly reflect the changes in the player's responses at each progressing levels. Here, unlike DAU based KPIs we explicitly track the level-induced changes by modeling the propagation in the playing characteristics as the gaming level increases. Figure 2 (a)-(c) show that player's engagement and achievement increase at higher levels of the game. We jointly model these drivers' of a player's level progressions as well as his dropout traits. Our model also accounts for the heterogeneity among players' behaviours. Section 5.1 describes the inter-dependencies among the players' engagement, achievement, collaboration and drop-out propensities.

In a single-player action mobile game, Banerjee et al. (2019) showed that joint analysis of player characteristics produces much improved marketing and general operations strategies. Here, we study players' responses in online RPGs which differ from single-player mobile games along multiple games. First, unlike Banerjee et al. (2019), here we have team-play and in-game collaboration among multiple players. These in-game player collaborations greatly influences game play volume. We estimate their social contagion effects in an action based RPG which is the case study in our data example (see section 5.3). Also, RPGs grant players an enhanced degree of control over their characters. Along with the development of direct gaming skills RPGs promote character growth through persistent exploration of the gaming world which is typically changed daily by the game developers. Combining an experience system into character development, action based RPGs seek to provide gamers the satisfaction of the action integrated with character growth and advancement.

To model how dropouts depend on the time-varying player motivations as the players progress through the game, we use a shared parameter model (SPM). SPMs have been very successfully used in longitudinal studies (Rizopoulos and Lesaffre, 2014, Rizopoulos et al., 2008, Vonesh et al., 2006) to address this bias from informative dropouts. Along with observed time-invariant covariates, we use correlated random effects to model the longitudinal outcome processes governing player's engagement, achievement and collaboration. In our SPM set-up, a player's dropout level depends on his motivations only through the random effects and observed covariates. Based on our estimated joint model, we provide level-wise predictions of future players' engagement, achievement, collaboration and dropout times.

For estimating JMLPC we use the methodology developed in Rizopoulos (2016) and Rizopoulos et al. (2009). Rizopoulos et al. (2009) developed a novel Laplace method for evaluating integrals associated with random effects in a joint modeling framework which is used in Rizopoulos (2016) for calculating the log-likelihood value for the posterior means of the parameters and the random effects, and for obtaining the marginal log-likelihood. Gaming data sets typically contain a large number of players and to achieve reasonable predictive power we need to incorporate player specific as well as level specific effects in our joint model. Modeling longitudinal data by a nonlinear cross-classified model with a large number of random coefficients can quickly become computationally intensive (Gao, 2017, Papaspiliopoulos et al., 2020, Zhang et al., 2016). Even in the comparatively simpler linear cross classified models, it follows from Gao et al. (2017) that likelihood-based inferences involving marginalisation over the factors are not scalable to gaming data sets with large number of players. Additionally, canonical Gibbs sampler is also not scalable in these applications due to its superlinear cost in the number of observations (Papaspiliopoulos et al., 2020). Our JMLPC framework uses multiple and heterogeneous longitudinal outcome variables and the shared parameters model allows for dependencies across the random effects of the outcomes. The methodology in Rizopoulos (2016) is not directly scalable for estimation in such framework with large number of players. To resolve this problem we undertake a *divide-and-recombine* (DR) approach. DR based techniques have been successful in providing scalable and accurate estimation in a wide range of regression problems (Battey et al., 2015, Chen and Xie, 2014, Jordan et al., 2019). In Section 4.1 using DR to estimate the joint model parameters, we scale the JMLPC modeling framework to accommodate the $10,000$ players present in our training data. In Section 6, we illustrate how our proposed DR based Bayesian estimation algorithm can be used to fit several variants and complex extensions of JMLPC.

3 Data

To improve a player's retention and lifetime engagement, we need to understand at what stage a player stops progression and drops out. We study players' motivations, progressions, and churn jointly from a lifetime span approach by randomly selecting $15,000$ players in an online action RPG who start from level one at day one. Their progressions in the following three months are observed. Our data include information on individual level-wise characteristics, social collaboration and game activities. The highest level reached by any player in this data set is 39.

Table 1 presents the summary statistics of these player characteristics for the different levels of the game. It shows the number of dropouts among the players who started each level. It provides the summary statistics values for engagement which is measured by the playing time (in seconds) needed to progress each level as well as that for achievement which is measured by points received for completing each level. Figure 2 shows the plots of these level-wise progression statistics. On average, engagement increases exponentially till level 9, then drops till level 13; between level 14 to 30, it increases with a very flat slope; and starting at level 30, the ascent is very pronounced. We observe 1.6% of the players progress to 39. Around 7.2% of these players used female game characters. Also, female characters progressed more as their representation went up from 7.2% in the lower level to around 35% in the higher levels.

Achievement scores showed similar trends on average (see subplots (a)-(c) of figure 2). The table and the figure show that there is a transition in gaming characteristics between levels 10 to 15. In Section 3, we use splines to capture this transition by allowing non-linear levels effects. The hazard rate also spiked between levels 11 to 15 reaching its peak value of 54% at level 14. Collaborations however steadily increased with level progression. It is to be noted that barring subplot (c), figure 2 shows statistics pertaining to the marginal distributions.

Geographically, the players were from 337 cities in China. We use corresponding city tiers (I-V) in our predictive framework. We follow the city tier system list published by China Business Network Co., Ltd. in 2019 which ranked 337 major

Table 1: Level-wise summary statistics aggregated across players.

Level	Started	Churned	Still	Hazard	Engagement (in sec)		Achievement Scores		Collaboration
No.	Level		Playing	Rate	Mean	Sd	Mean	Sd	Proportion
1	15000	10	2	0.07%	205.4	160.0	5963.1	846.7	2.2%
2	14988	16	0	0.11%	559.8	420.6	11932.6	1810.4	2.8%
3	14972	28	1	0.19%	1065.7	984.5	20050.8	4836.5	3.2%
4	14943	69	2	0.46%	1875.1	1637.2	31472.7	10109.1	3.4%
5	14872	63	3	0.42%	3093.4	2547.4	53239.8	28873.7	3.7%
6	14806	62	5	0.42%	3692.4	2984.4	61499.7	33534.3	3.9%
7	14739	229	5	1.55%	4210.7	3340.1	70321.0	37896.9	4.1%
8	14505	640	11	4.41%	4958.4	3849.9	84108.7	45947.8	4.3%
9	13854	1599	32	11.54%	5330.4	5029.3	80206.7	53880.5	5.6%
10	12223	1222	24	10.00%	4803.1	5706.8	85795.6	69497.6	6.7%
11	10977	2513	44	22.89%	5273.3	5402.3	108225.8	90372.3	8.5%
12	8420	2276	35	27.03%	4306.7	5521.9	76972.8	59611.3	11.5%
13	6109	1288	17	21.08%	3745.6	5349.6	67191.6	56640.1	14.1%
14	4804	2627	44	54.68%	5197.3	7939.5	55138.8	44587.5	28.4%
15	2133	470	36	22.03%	5195.6	7634.2	51717.9	41599.5	34.3%
16	1627	223	28	13.71%	6322.8	11837.5	60484.5	51492.8	40.0%
17	1376	97	18	7.05%	6348.8	10770.5	62832.8	58386.9	42.8%
18	1261	91	31	7.22%	6192.4	7661.0	67439.9	75542.9	44.5%
19	1139	53	23	4.65%	5680.4	7089.7	67548.0	63753.9	44.3%
20	1063	33	16	3.10%	5980.3	7733.7	71187.6	65278.2	45.0%
21	1014	22	15	2.17%	5901.0	7368.0	68375.9	59900.5	45.0%
22	977	28	31	2.87%	6225.4	6940.4	70427.9	64969.1	47.5%
23	918	74	55	8.06%	6897.6	7832.5	74320.6	74343.7	52.9%
24	789	27	57	3.42%	7256.3	8999.9	80392.8	64899.4	52.8%
25	705	21	45	2.98%	8017.0	10789.8	81159.3	64293.0	54.8%
26	639	12	39	1.88%	7750.0	9092.0	83416.7	67470.5	59.9%
27	588	13	43	2.21%	7820.0	8778.5	85300.2	70752.9	63.5%
28	532	5	67	0.94%	8791.3	11981.3	79931.4	71697.7	62.0%
29	460	7	94	1.52%	9832.4	10431.4	85983.5	76434.9	77.7%
30	359	1	29	0.28%	11145.3	11194.6	93578.8	102253.4	83.6%
31	329	3	56	0.91%	14894.5	14134.2	117095.5	111583.7	90.0%
32	270	0	3	0.00%	13263.8	14098.0	172299.8	208628.3	91.8%
33	267	2	2	0.75%	13124.9	12052.5	178823.0	205752.8	92.4%
34	263	2	2	0.76%	11745.0	11780.9	174256.1	170769.7	91.5%
35	259	0	1	0.00%	13212.8	15214.3	188498.3	171717.9	93.0%
36	258	1	3	0.39%	15321.9	15547.8	212545.0	206667.9	93.7%
37	254	1	3	0.39%	15963.6	15881.3	209800.0	189389.8	91.2%
38	250	0	4	0.00%	16087.6	16994.8	215173.9	194288.9	88.2%
39	246	17	229	6.91%	19190.5	20455.7	244247.9	247657.3	96.7%

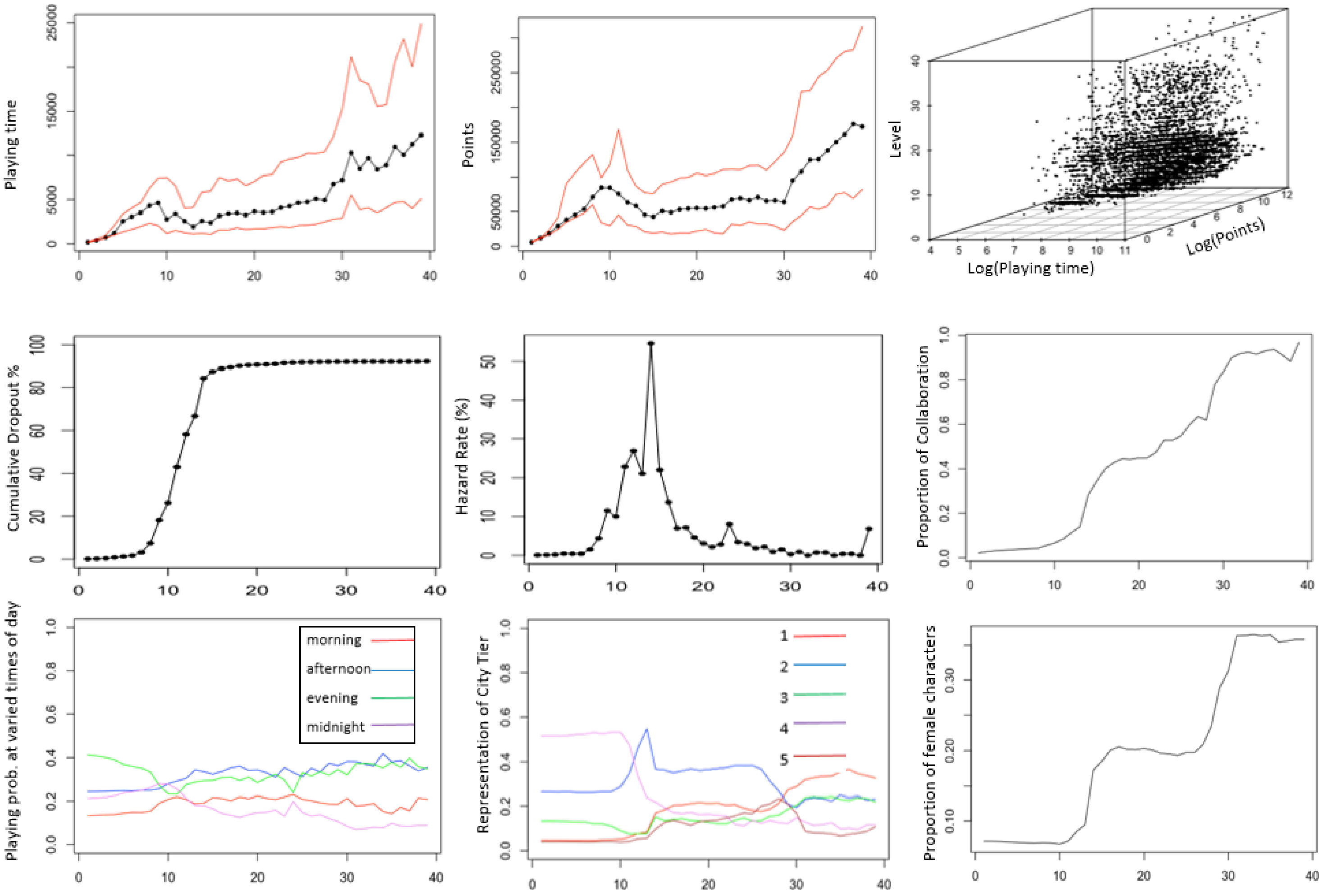
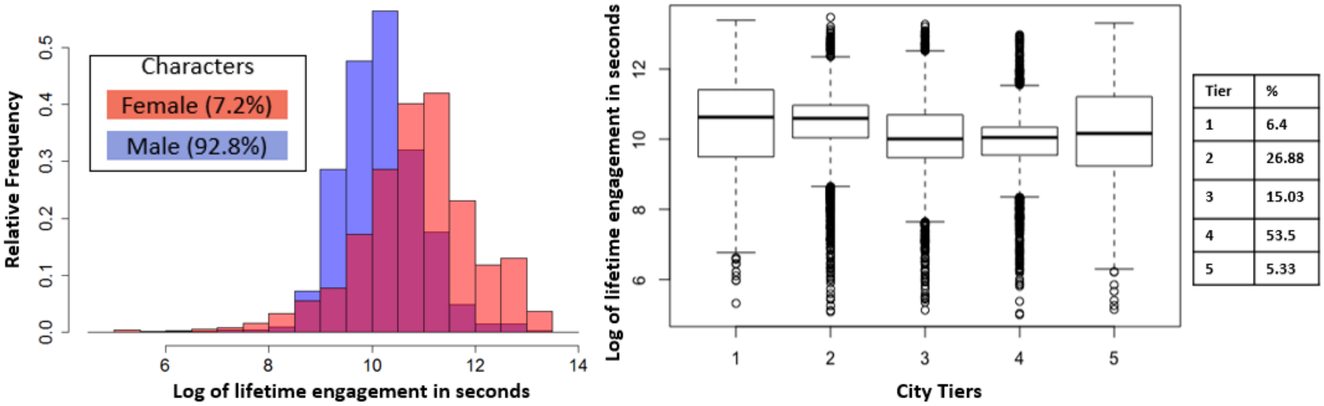


Figure 2: Level-wise progression statistics aggregated across players. Clockwise from top-left we have (a) Engagement as measured by the playing time (in seconds) needed to progress the level (the median line is in black and the quartiles are in red lines) (b) Achievement scores measured by points received for completing each level (the median line is in black and the quartiles are in red) (c) 3D plot of logarithm of playing time (engagement), points (achievement) and level (d) Cumulative churn rate (e) Hazard rate (f) Collaboration proportion across levels (g) Distribution of playing times in the level according to the different times of day (h) Proportion of players from different city tiers who completed the level (i) Proportion of female game characters that completed the level. All the plots except (c), are aggregated across players and have level plotted along the x-axis.

cities into six categories based on economic activities. We merged the last two categories to category V. Out of the $15,000$ players $1,155$ were still active in the game and so their dropout levels and lifetime game trajectory are missing. Figure 3 shows the distribution of players' lifetime engagement in the game across character genders and city tiers. For each level, a player's playing time decomposed at four different time periods of the day (morning, afternoon, evening, midnight) were recorded. On average, afternoon was the most popular time slot. Along with these information, we also have level-wise progression records of players in-game engagement, collaboration and achievement and the stage where he dropped out (unless he/she was an active player).



Tier	%
1	6.4
2	26.88
3	15.03
4	53.5
5	5.33

Figure 3: Distribution of retention times: (a) histogram by gender of game characters (b) box plots by city tiers.

We study the inter-dependencies between these attributes by modeling their joint distribution in the following sections. We use $10,000$ players for training our proposed joint model and $5,000$ players are used as test set and validation purposes.

4 Joint Modeling of Level-wise Playing Characteristics: JMLPC

For the kth player E_{lk}, C_{lk} and A_{lk} record his/her engagement, collaboration and achievement at the l th level where $l = 1,\dots,L_k$ and $k = 1,\dots,n$ with $n = 10000$; L_k is the maximum level that the kth player has played till date. The binary variable D_k records if the player has dropped out (1) or is still playing (0). Thus, if $D_k = 0$, L_k is the current level that the kth player is playing. We next describe a simple joint model JMLPC for modeling the different level-wise playing characteristics of a gamer. Our approach is fairly general and can encompass several variants and extensions of the model. We describe the simple JMLPC framework below and then discuss the extensions in Section 6.

We model Engagement (in secs) and Achievement (in raw scores) by a log-linear model with level (l) dependent intercepts α_l as well as player (k) specific coefficients $\beta_{kl}^{(i)}$ as follows:

$$\log E_{kl} = \alpha_l^{(1)} + \delta^{(1)}\,\mathrm{F}_k + \beta_{kl}^{(1)}\ , \tag{1}$$

$$\log A_{kl} = \alpha_l^{(2)} + \delta^{(2)}\,\mathrm{F}_k + \beta_{kl}^{(2)}\ . \tag{2}$$

F_k is a dummy variable representing player k character's gender with 1 denoting female. Thus, $\delta^{(1)}$ and $\delta^{(2)}$ are level-independent exponential changes in engagement and achievement of a female over a male character. The model in (1) is not identifiable. Later, we impose further constraints on α and β which would make the model identifiable.

We model the binary collaboration indicator C_{kl} by a logistic regression as:

$$\texttt{logit}(P(C_{kl} = 1)) = \alpha_l^{(3)} + \delta^{(3)}\,\mathrm{F}_k + \sum_{j \in J} \gamma_j\,\mathrm{T}_{jkl} + \beta_{kl}^{(3)}\,. \tag{3}$$

In our data set, only approximately 10% of C_{kl}s are positive and so, we refrained from considering finer details on the nature of collaborations such as the number of collaborators and their expertise (current level) and concentrated solely on the indicator variable for measuring collaboration. Apart from gender of the game character, the proportion of time T_{jkl} that player k played in day part j at level l, is also used as a covariate in (3) as platform load might fluctuate at different parts of the day. Here, $J = \{\text{morning}, \text{afternoon}, \text{evening}\}$. We have used time of the day in (3) and not in (1), (2) as we are more interested in observing if the load in the platform affects collaborative activities. Ideally, all the control variables should be considered in (1)-(3) which is done later in Section 6.

The player invariant intercepts $\{\alpha_l^{(i)} : l = 1, \ldots, 39\}$ captures the varying gaming difficulty across levels. Figure 2 displays the smoothness and non-linear growth patterns in the average statistics associated with engagement, achievement and collaboration. Thus, we impose smoothness constraints on them. We use natural cubic splines f_i with κ_f knots for modeling the player invariant intercepts, i.e.,

$$\alpha_l^{(i)} = f_i(l; \boldsymbol{a}^{(i)}) \text{ for } l = 1, \ldots, 39 \text{ and } i = 1, 2, 3. \tag{4}$$

This imposes level-based contiguity on the effects allowing for non-linear growth patterns. The spline coefficients $\boldsymbol{a}^{(i)}, i = 1, 2, 3$ are estimated in the joint modeling framework. Later in tables 6 and 7 we discuss the implications of these smoothness constraints by comparing with a joint model without such constraints.

Next, we impose level-based smoothness on the player specific effects $\beta_{kl}^{(i)}$. Consider

$$\beta_{kl}^{(i)} = s_i(l;\, \boldsymbol{b}_k^{(i)}) \text{ for } l = 1, \ldots, 39;\ k = 1, \ldots, n \text{ and } i = 1, 2, 3; \tag{5}$$

where, s_i is a natural cubic spline on levels l with κ_s knots whose canonical basis coefficients are given by $\boldsymbol{b}_k^{(i)}$. So, the form of the spline function s_i is invariant across players but the coefficients (and hence the shape) vary across players for $k = 1, \ldots, n$. With these natural level-based smoothness constraints, the model (1)-(5) is identifiable.

We impose a second-level hierarchical structure on the player specific effects by considering a correlated (across i) prior structure on their basis coefficients. We consider $\boldsymbol{b}_k = (b_k^{(1)}, b_k^{(2)}, b_k^{(3)})$ to be independent and identically distributed across players $k = 1, \ldots, n$ from a Gaussian distribution with mean zero, i.e., $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$. The correlation structure Σ

is estimated based on the data and captures the inter-dependencies in a player's engagement, achievement, and collaborative actions.

Next, we model the dropout probabilities. Define the hazard rate of player k for dropping out at level l as $\lambda_{kl} = P(D_k = 1 \mid L_k = l)$. Then, the dropout probability of player k is: $P(D_k = 1) = \lambda_{kL_k} \prod_{l=1}^{L_k - 1}(1 - \lambda_{kl})$. We estimate the hazard rate through a shared parameter model:

$$\text{logit}(\lambda_{kl}) = \mu_l + \nu_1 \text{F}_k + \sum_{i=1}^{4} \nu_{2,i} R_{ki} + \eta_1^T \, \boldsymbol{b}_k^{(1)} + \eta_2^T \, \boldsymbol{b}_k^{(2)} + \eta_3 \, C_{k,l-1} \tag{6}$$

where, R_{ki}s are dummy variables corresponding to the five different city tiers is and μ_l is the level specific intercept. In the presence of these level specific intercepts μ_l in (6), we apprehend that residual player specific effects of engagement and achievement (above or below the level average) will be impacting the hazard rate. For example a highly motivated player would be involved in more explorations and quests than average and it would be reflected in his over average engagement and achievement. On the contrary, larger raw engagement and achievement values do not necessarily indicate player enthusiasm as those vary across the level of the game (as seen in table 1). Thus, for estimating the hazard rate in our shared effect framework in (6), we consider the residual player specific effects for engagement and achievement and not their current values. As collaboration $C_{k\cdot}$ is an indicator variable, we used its value in the penultimate level in (6). Next, we describe estimation of model in (1)-(6).

4.1 Estimation: Implementation and Scalability

For writing the likelihood function for our model let $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n)$, and Θ be the collection of all the other coefficients. Our model assumes that the parameters $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$, account for all the dependencies of between the observed data. Thus, conditioning on them the outcomes, engagement, achievement and collaboration indicators, as well as their measurements for each level are independent of each other, and of players' dropout. Thus, the complete likelihood $\text{CL}(\Theta, B)$ equals $\prod_{k=1}^{n} \mathcal{L}_k(\Theta, \boldsymbol{b}_k)$ with $\mathcal{L}_k(\Theta, \boldsymbol{b}_k)$ being proportional to

$$\lambda_{kL_k}^{D_k} \prod_{l=1}^{L_k - 1} (1 - \lambda_{kl}) \prod_{l=1}^{L_k} \text{PE}_{kl} \, \text{PA}_{kl} \, \text{PC}_{kl} \; ,$$

where, λ_{kl} is given by (6) and

$$\begin{aligned}
\text{PE}_{kl} &= P(E_{kl} \mid \boldsymbol{a}^{(1)}, \delta^{(1)}, \sigma^{(1)}, \boldsymbol{b}_k^{(1)}) = \phi\big(\log E_{kl} - f_1(l; \boldsymbol{a}^{(1)}) - \delta^{(1)} \text{F}_k - s_1(l; \boldsymbol{b}^{(1)});\, \sigma^{(1)}\big), \\
\text{PA}_{kl} &= P(A_{kl} \mid \boldsymbol{a}^{(2)}, \delta^{(2)}, \sigma^{(2)}, \boldsymbol{b}_k^{(2)}) = \phi\big(\log A_{kl} - f_2(l; \boldsymbol{a}^{(2)}) - \delta^{(2)} \text{F}_k - s_2(l; \boldsymbol{b}^{(2)});\, \sigma^{(2)}\big), \\
\text{PC}_{kl} &= P(C_{kl} \mid \boldsymbol{a}^{(3)}, \delta^{(3)}, \boldsymbol{\gamma}, \boldsymbol{b}_k^{(3)}) = \frac{\exp[C_{kl} \cdot (f_3(l; \boldsymbol{a}^{(3)}) + \delta^{(3)} \text{F}_k + \sum_{j \in J} \gamma_j \, \text{T}_{jkl} + s_3(l; \boldsymbol{b}^{(3)}))\,]}{1 + \exp\big(f_3(l; \boldsymbol{a}^{(3)}) + \delta^{(3)} \text{F}_k + \sum_{j \in J} \gamma_j \, \text{T}_{jkl} + s_3(l; \boldsymbol{b}^{(3)})\,\big)},
\end{aligned}$$

with $\phi(\,.\,;\sigma)$ denoting univariate normal density with mean 0 and standard deviation σ. We follow a Bayesian estimation procedure. We use the following prior specifications. For the parameters $\boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\nu}, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2$ and η_3 we use independent components of normal prior with zero mean and variance 1000 (which makes the prior nearly non-informative). We use independent inverse Gamma priors for the variance parameters $\sigma^{(1)}$ and $\sigma^{(2)}$.

For both s_i and f_i, $i = 1, 2, 3$, we consider natural cubic splines with two knots at $l = 9$ and 15; this choice was based by assessing the empirical transitions in the median engagement and achievements across levels as witnessed in Figure 3 and described in Section 2. On the basis coefficients $(\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, \boldsymbol{a}^{(3)})$, non-informative product Gaussian prior with mean 0 and variance 1000 was used. On the player specific basis coefficients $\boldsymbol{b}_k$, the prior used was $N_{12}(0, \Sigma)$. The covariance matrix Σ was parametrized with the correlation matrix and the vector of variances. We used Lewandowski-Kurowicka-Joe (LKJ) prior distribution for the correlation matrix and product half t-distributions for the variance vector. The estimates of $\boldsymbol{\mu} = \{\mu_l : l = 1, \ldots, 39\}$ were based on natural cubic splines with knots adaptively selected based on the data.

The above estimation procedure is implemented by using the R package of Rizopoulos (2016) which uses Hamiltonian Monte Carlo (Neal et al., 2011). However, as we had player specific random effects in the glmm set-ups of (1)-(6), the package was not readily scalable to accommodate the very large scale glmm estimation problem that we encounter with our training data-set of $10,000$ players. Scaling mixed effects models in massive *id-level* data for modeling user's personal preferences to a product is an important topic in current statistical research (see Zhang et al., 2016 and the reference therein). To address this big-data problem, we use a divide-and-recombine approach. Divide and recombine (D&R) approaches (Jordan, 2012) have been witnessed to provide accurate estimates in big-data regression problems (Battey et al., 2015, Chen and Xie, 2014, Jordan et al., 2019). Modifying the methodology of Rizopoulos (2016) in a D&R big-data set-up, we conduct estimation of model (1)-(6) by an iterative algorithm.

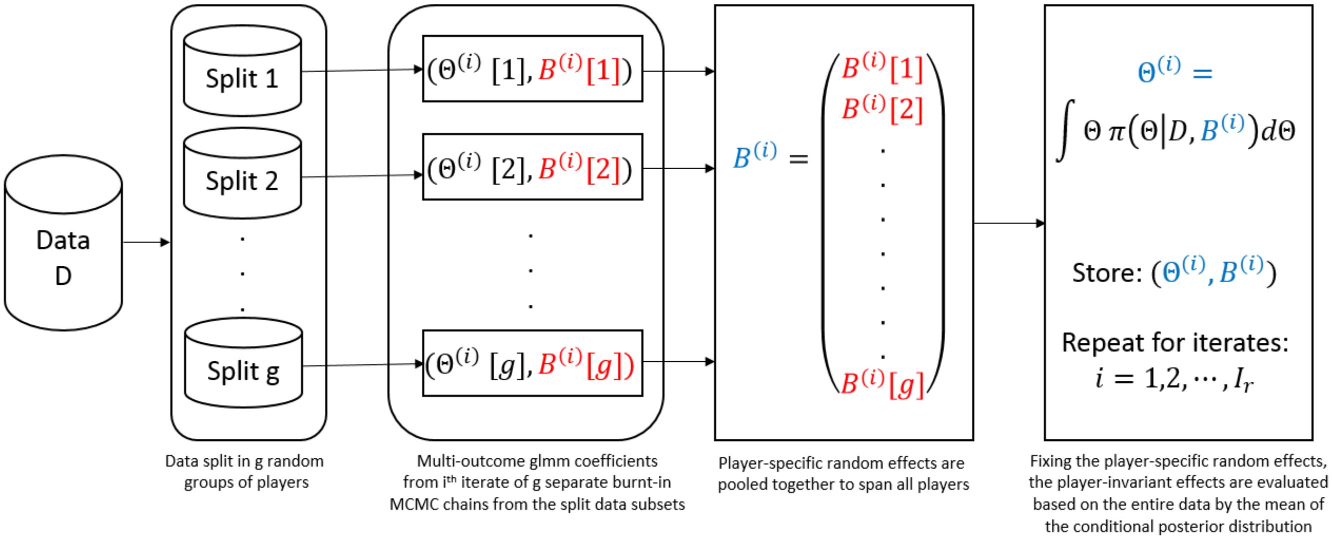

Data
D
Split 1
Split 2
Split g
$(\Theta^{(i)}[1], B^{(i)}[1])$
$(\Theta^{(i)}[2], B^{(i)}[2])$
$(\Theta^{(i)}[g], B^{(i)}[g])$
$B^{(i)} = \begin{pmatrix} B^{(i)}[1] \\ B^{(i)}[2] \\ \vdots \\ B^{(i)}[g] \end{pmatrix}$
$\Theta^{(i)} = \int \Theta\, \pi(\Theta | D, B^{(i)}) d\Theta$
Store: $(\Theta^{(i)}, B^{(i)})$
Repeat for iterates: $i = 1,2,\cdots,I_r$
Data split in g random groups of players
Multi-outcome glmm coefficients from ith iterate of g separate burnt-in MCMC chains from the split data subsets
Player-specific random effects are pooled together to span all players
Fixing the player-specific random effects, the player-invariant effects are evaluated based on the entire data by the mean of the conditional posterior distribution

Figure 4: Schematic diagram of the *divide & recombine* joint modeling algorithm.

A schematic diagram of the proposed D&R joint modeling algorithm is presented in Figure 4. The iterative D&R JMLPC algorithm is as follows:

1. Split $n = 10,000$ players into $g = 20$ groups of $m = 500$ players each.

2. Estimate the multi-outcome mixed effects model (1)-(5) on each of the g splits separately. After burn-in, $I_r = 1000$ successive iterations from each of the g chains are kept.

3. Pool the estimates of the mixed effects from the g splits and reestimate the fixed effects as follows. For each $i = 1, \ldots, I_r$:

   a. Pool the i^{th} stored iterate from the g splits and construct $B^{(i)}$ as the pooled list of all estimated random effects coefficients.

   b. Keeping the random effects $B^{(i)}$ fixed, run MCMC for the fixed effects for each outcome variable separately and store $I_f = 1000$ iterates from chain of fixed effects $\{\Theta^{(ij)} : j = 1, \ldots, I_f\}$ after burn-in.

   c. The re-estimated fixed effects coefficients $\Theta^{(i)}$ are the posterior mean from the corresponding chain $\{\Theta^{(ij)} : j = 1, \ldots, I_f\}$.

4. Estimate the joint model using the chain for the longitudinal mixed effects model $\{(B^{(i)}, \Theta^{(i)}) : i = 1, \ldots, I_r\}$ and using `mvJointModelBayes` function in `JMbayes` package of Rizopoulos (2016). This conducts joint estimation of the hazard rates along with the longitudinal outcomes.

5 Results

5.1 Fitted Model: Coefficients and Interpretation

Our joint model was estimated on 10,000 players playing a total of 123,681 game levels using the algorithm described in Section 4.1, and illustrated in Figure 4.

For the longitudinal variables, Figure 5 and Figure 6 look at estimation of the random effects $B^{(i)}$'s and the fixed effects $\Theta^{(i)}$'s respectively. The random effects are estimated in several splits of the data and then pooled. The left plot of Figure 5 compares the random effects estimated from the split of the data that was used (the red lines) to other random splits of the data (the reference boxplots). To create the reference boxplots we randomly split our 10,000 players into 20 parts many times, and from each of these random splits of 500 players we estimated the Frobenius norm of the random coefficients corresponding to the three outcome variables. The right plot in Figure 5 shows the estimated correlation matrix of 12 player specific random coefficients B. Generally, the random effects are positively correlated, not only for each of the outcome variables, but also across three outcome variables. This indicates that a player who has a greater engagement in the game-play is also more likely to have higher achievement and more collaboration.

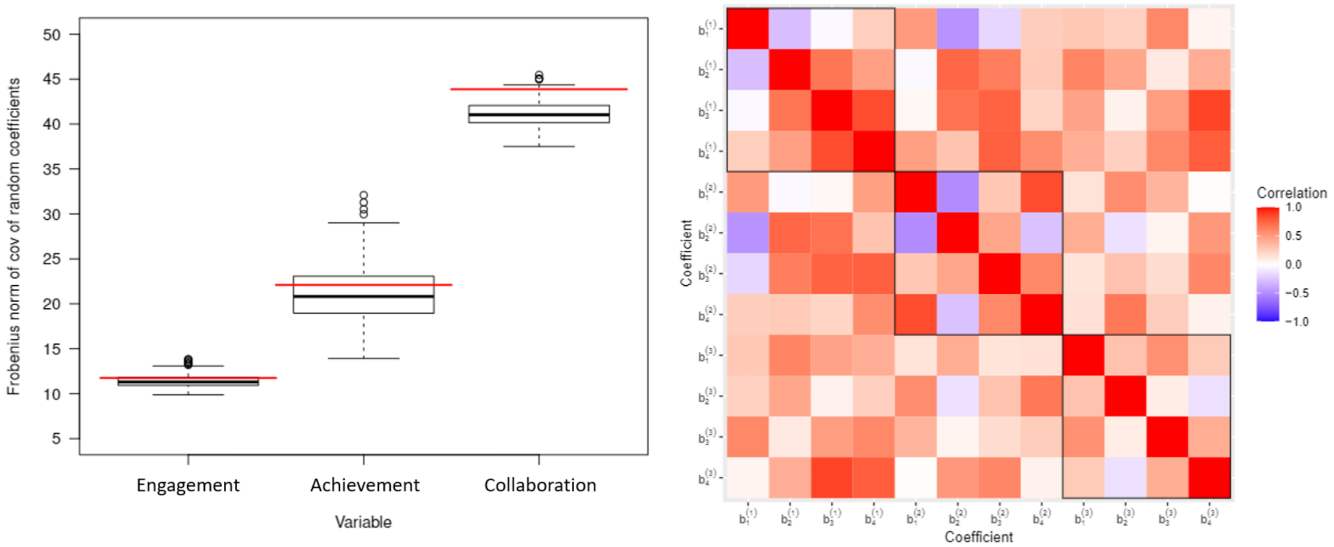


Figure 5: Posterior estimate of the random effects in models (1)–(3). The left plot shows in red lines the average of the 20 Frobenius norm of the covariance of the random effects of the players in each split. The reference boxplots are from 25 different random splits of the 10000 players in 20 groups. The right plot shows the posterior correlation matrix of the all 12 random effects.

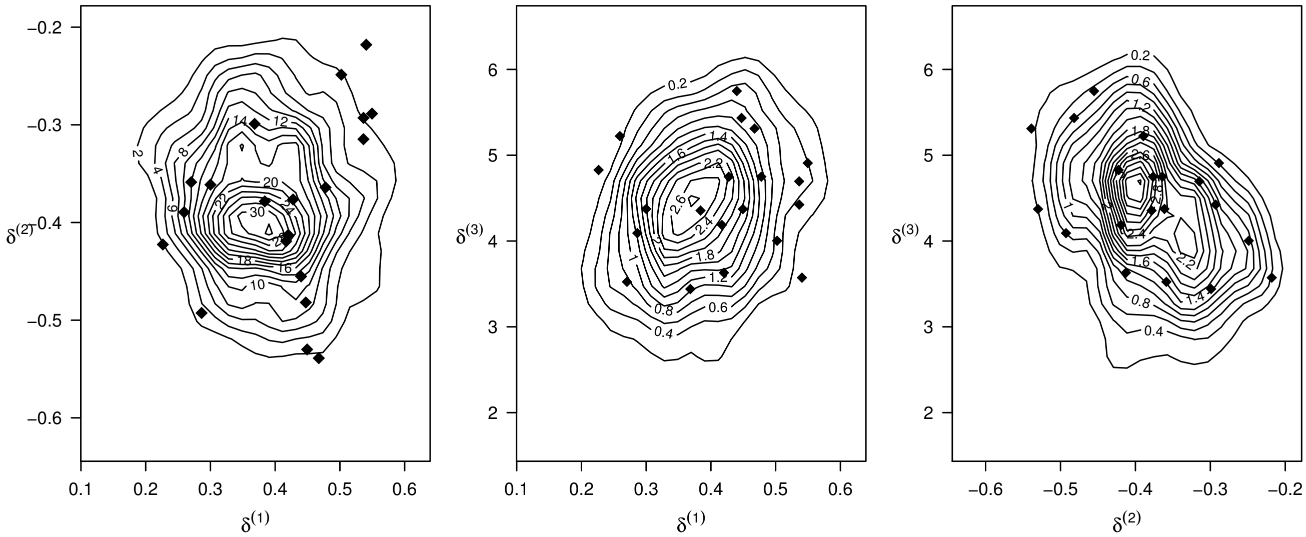


Figure 6: Posterior estimate of the parameters of the outcome models (1)–(3). The contour plots are from the reestimated posterior of the fixed effects. The points in the plots are the means of the estimated distribution from the 20 separate splits of the data, before reestimation.

During our estimation of the random effects, we also had an estimate of the fixed effects from different splits of the data; these are denoted by $\Theta^{(i)}[1], \dots, \Theta^{(i)}[20]$ in Figure 4. But since these estimates only use the players in those splits, we re-estimated these fixed effects in another step that runs a MCMC conditioning on the pooled random effects estimates. Figure 6 provides a comparison of our fixed effects estimates to their previous estimates. The 20 points in each plot show the posterior mean of the parameters estimated in 20 splits. Clearly, the estimates vary in different splits where a smaller number of players are used. The contour plots are of the posterior distribution of the fixed effects parameters after re-estimation that uses all the players. In the first plot of Figure 6, we see that sometimes the estimates from a split was far from the our final estimate. Additionally, the spread of the parameter estimates in the splits were higher than the re-estimated posterior spread. For example, the average over the splits of the estimated standard deviations for $\delta^{(3)}$ was 0.82, and the standard deviation of $\delta^{(3)}$ from the re-estimated chain was 0.73.

Coefficients for gender in the outcome models are statistically significant. They are significantly positive for engagement and collaboration, but significantly negative for achievement. Note that, gender F_k was coded as 0 for player k playing with male character and 1 for player k playing with female character. Thus, there is higher level of engagement and collaboration for female characters compared to male characters, but a lower level of achievement. Specifically, we estimate on average 47% more engagement (95% credible interval, 27% to 72%) in game-play for players playing with a female character compared to a player playing with a male character. But, a female character gets on average 68% (95% credible interval, 60% to 78%) of achievement points compared to what a male character achieves in similar situation. We also estimate that playing during the afternoon had a higher propensity to play in teams than if for playing during the morning. The odds ratio of playing in a team in the afternoon is 1.53 (with 95% credible interval, 1.12 to 2.08) relative to in the morning. Actually, gameplay in the afternoon showed the highest level of collaboration than any of time of the day.

Table 2: Posterior estimate of the parameters the hazard model in the joint model (1)–(6)

Parameter	Posterior mean	95% Credible interval
Gender (0 = male, 1 = female)		
ν_1	-0.5957	(-0.7484, -0.4412)
Engagement		
η_{11}	-1.015	(-1.7598, -0.3613)
η_{12}	-0.1229	(-0.1982, -0.0499)
η_{13}	-0.2076	(-0.5022, 0.0652)
η_{14}	0.1705	(-0.0753, 0.3836)
Achievement		
η_{21}	-1.247	(-3.7933, 1.2143)
η_{22}	0.0092	(-0.0828, 0.0972)
η_{23}	-0.0708	(-0.2783, 0.1668)
η_{24}	-0.0013	(-0.0795, 0.0717)
Collaboration		
η_3	-0.0285	(-0.0505, -0.0091)

The estimates of several of the parameters in the hazard model in equation (6) are given in Table 2. The posterior estimate shows that player with a male character has $\exp(-\nu_1) = \exp(0.5957) = 1.81$ times (95% Credible interval, 1.55 to 2.11) more chance of dropout of the game compared to female character. Collaboration is negatively associated with dropout. Accounting for the logistic model for collaboration in equation (3) the joint model estimates that playing in team decreases the dropout probability by 1.03 times (95% Credible interval, 1.01 to 1.05). The coefficients in the hazard model (6) that connects dropout to engagement modeled in (1) are $\boldsymbol{\eta}_1 = (\eta_{11}, \ldots, \eta_{14})^\top$. The corresponding posterior means of these parameters are negative; except for η_{14} whose credible interval contains 0. Thus, for example, a player with better than average achievement has a lower chance of dropout. But the amount by which the chance is lowered also vary by the level.

5.2 Retention Profiles of Future Players

Predicting the retention profiles of future players is very important for managerial research. Such prescriptive analysis of new player behavior is fundamental for the maintenance of existing as well as for the creation of new advertisement based monetization routes in these gaming platforms. Using our proposed model we predict the retention probabilities of future players at the different levels of the game as well as their cumulative playing time to reach each those levels. Table 9 presents the level-wise summaries of the average retention probabilities as well as the average cumulative playing time across future players from different city-tiers who would be playing with male and female characters. The expected lifetime engagement of a future player was computed as ELE $= \sum_{l=1}^{L} \hat{p}_l \, \hat{E}_l$ where $\hat{p}_l$ and $\hat{E}_l$ are the predicted average retention probabilities and average playing times for level l. The predictive distribution of the LE for a random future player was evaluated. Figures 7 and 8 depict these predictive traits. We see that players playing with female characters have higher life-time engagement. Also, players from Tier I cities have the highest retention probabilities. The expected life-time engagement is lowest for Tier III cities. As such, players playing with male characters (who constitute more than 90% of the platform) has future life-time engagement values 46.6%, 33.4%, 15.8% and 19.7% higher for tier I, II, IV and V cites respectively compared to tier III cities. If player acquisition costs are similar across different tiers, it will benefit to market the game more in tier I cities for attracting new players and increase representation from tier I cities.

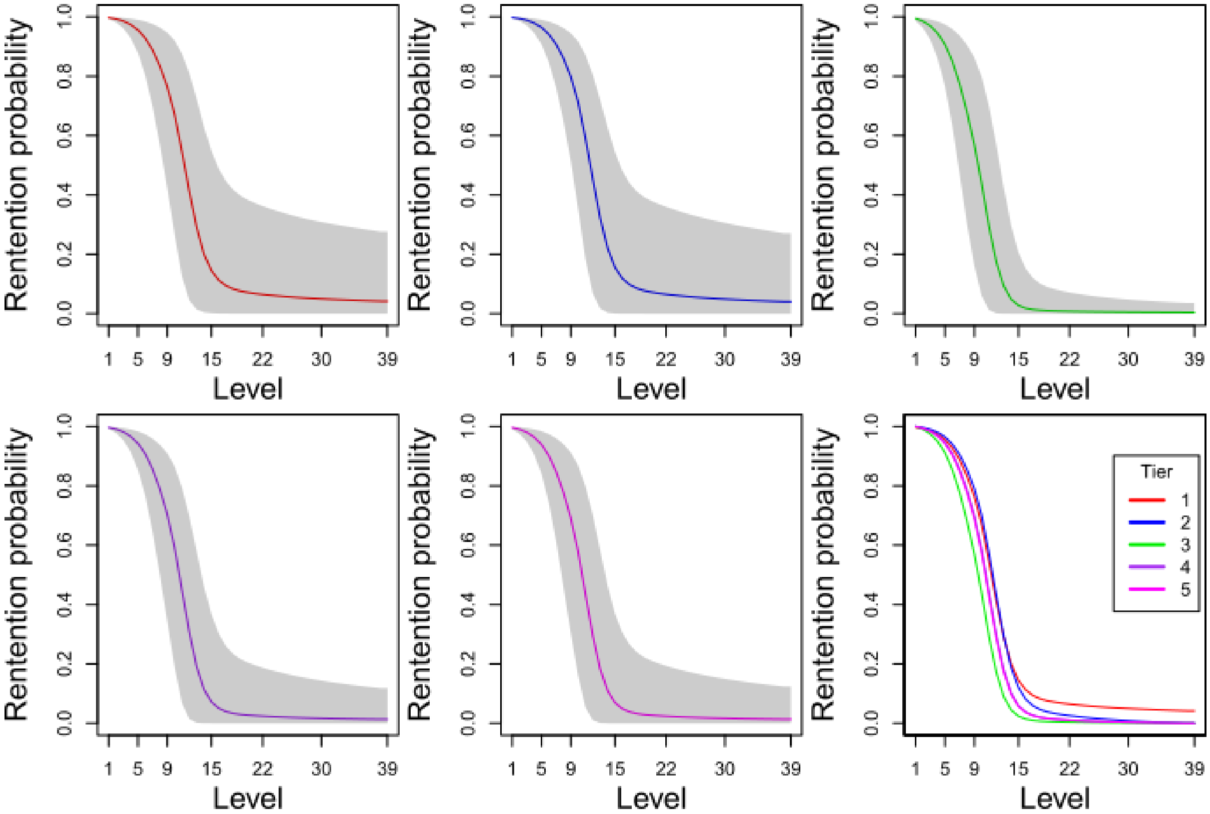


Figure 7: Expected retention probabilities of future players are plotted for different city tiers in the bottom right subplot. The 95% prediction interval for retention probabilities is shaded along with their expected values for city tiers I to V row wise starting from top left.

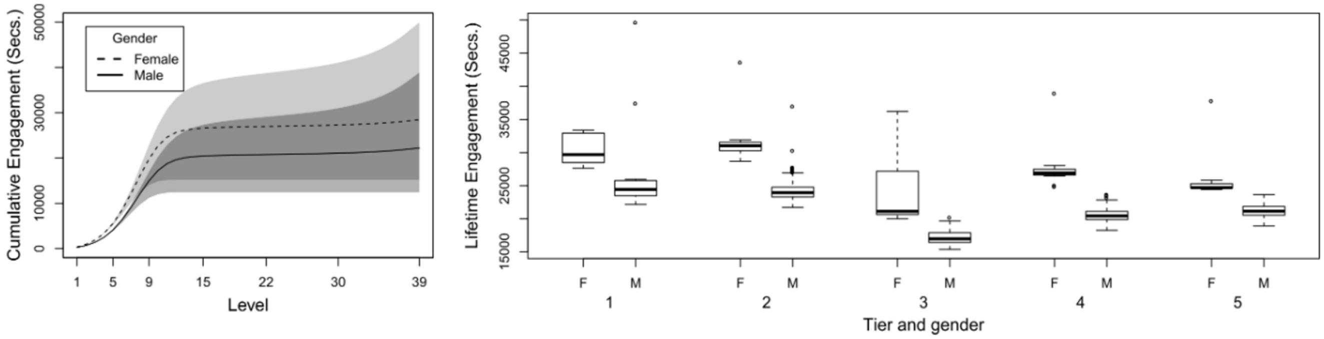


Figure 8: In the left panel, the expected cumulative engagement is plotted across levels in dotted and continuous lines for female and male characters respectively. The 95% prediction interval is also plotted. In the right panel, the box plots show the predictive distributions of lifetime engagements across city tiers and character gender.

5.3 A Marketing Retention Application

Collaboration is an important attribute in online RPGs. Managers can regulate collaboration efforts in these platforms by providing in-game incentives for increased team work. Incentives such as enhanced weaponry, extra character life or virtual currency that can be used for purchasing improved game amenities are well used in online games for retention marketing and promotions. Modern managers need to predict the effects of promotions of different magnitudes so that he/she can conduct field experiments and higher level validation on a short list of selected optimal choices. Simulation based on our prescribed joint model account for the randomness in the players' decision-making and can be used for furnishing the manager with a range of possible outcomes, and their probabilities of occurrence for each possible course of managerial actions. We demonstrate one such application below.

Consider two scenarios based on a mild promotion and an aggressive promotion campaign that increase the odds for collaborating respectively by 15% and 40% uniformly across the game. Simulating player behaviours in these scenarios can be easily done based on our estimated JMLPC by changing the estimates for $\alpha_l^{(3)}$ in (6) and keeping all other estimates in (1)-(6) invariant. Figure 9 shows how the predictive distributions of the lifetime player engagement values changes over the baseline due to applications of mild and heavy collaboration inducing promotions. Utilizing our model formulation, we calculate the benefits of promotion on engagement at different stages of the game. The expected engagement which is the product of the expected retention probability and the expected playing time of retained players, attains its maximum at level 8 in the baseline and mild encouragement cases and at level 9 for high encouragement case. This is due to that fact that the retention probabilities sharply decrease in those levels. In the baseline model, the minimum value of expected engagements is attained at level 22 after which average engagement increases with levels; the ascent is particularly steep after level 32. At level 22, mild promotion is witnessed to produce 44% increased engagement whereas aggressive promotions increase average engagement by 143%. Table 10 shows the level-wise changes due to application of the collaboration inducing promotions. The expected lifetime engagement value of a future player is seen to increase by 6.9% over baseline for mild and 20% for aggressive promotions.

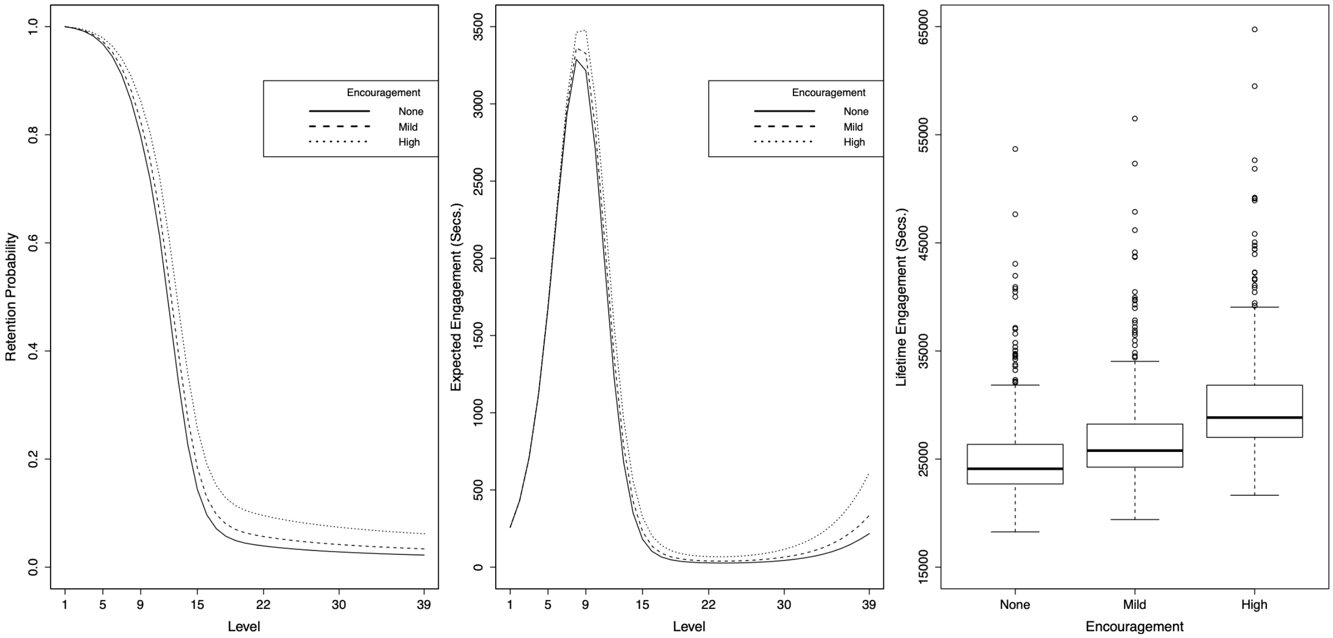


Figure 9: Plots of expected retention probability (ERP) and expected engagement (which is ERP times expected playing time of retained players) for a random player as mild and high collaboration encouragements are introduced in the gaming portal. The right most plot demonstrates the predictive distribution of lifetime engagement.

In table 10, we also report the value of a player based on premium add-on components purchase. In PP (premium purchase) value column of table 10, the average revenue that the gaming company makes per player is reported for unit playing time in each level. We did not have player specific purchase information but we had access to the portal wide average statistics for our RPG game. Using this portal average statistics, we calculate the expected premium purchase value (EPPV) for a future player based on the predictions from our estimated model. Unlike lifetime engagement, whose monetization needs planning and placements of advertisements, EPPV corresponds to the direct revenue that can be obtained from a future player. Table 10 provides EPPV for the baseline model as well as for mild and aggressive collaboration encouragements. The expected lifetime PPV of a future player in the baseline, mild and aggressive promotion model are ¥1.40, ¥1.53 and ¥1.79 respectively. Thus, collaboration encouragement can produce 27.4% increase in direct revenue.

5.4 Predictive Performance

We compare the predictive performance of our method to following two approaches that are widely used in retention marketing. The first one is perhaps the most naive and also the most popular. It uses the empirical distributions of engagement, achievement, collaboration and dropout proportions based on the training set to predict attributes of players in the test set. Here, the inter-dependencies between player motivations and level progressions are not used in predicting dropout probabilities. At each particular level, the marginal distributions of engagement, achievement, collaboration and dropout proportions from the training data are used for predicting the responses in the test data. The second approach uses *Cox Proportional Hazards Model*

(CPHM) (Cox, 1972, Efron, 1977) estimated on the training set for predicting dropout probabilities in the test set. CPHM models the dropout hazard λ_{kl} for player k at level l as $\lambda_{kl} = \lambda_{0l}\exp(\theta_f F_k + \sum_{j\in J}\theta_{t,j}T_{jkl} + \sum_{i=1}^{4}\theta_{r,i}R_{ik} + \theta_c C_{k,l-1})$, where λ_{0l} is the baseline hazard. The θ coefficients are estimated on the training data. Unlike the empirical approach, in CPHM a player's engagement, achievement and collaboration record at the level is used for predicting dropout. However, it does not use inter-dependencies between the player motivation metrics that is used in our joint modeling set-up. To compare the predictive performance of these methods, we consider all players in the test set who completed level 10 and drop-out of the game at one of the succeeding levels. We denote the set of such players by $\mathcal{T}$. Using their in-game history up to level 10 in our estimated joint model of (1)-(6), we predict their engagement times, achievement scores, collaboration and dropout probabilities at the future levels of the game.

Table 3: Retention probability prediction at future levels using the estimated JMLPC model and gaming history till level 10. We report the average and the 95% interval of the predicted retention probabilities in RP.

Level		RP	OR	MAD
20	Truth	0.06458	-	-
	JMLPC	0.07491 (0.00326, 0.14615)	2.37175	0.12001
	CPHM	0.08184 (0.02276, 0.14087)	1.81832	0.12915
30	Truth	0.02351	-	-
	JMLPC	0.04562 (0.00133, 0.07994)	3.76030	0.04562
	CPHM	0.06047 (0.00537, 0.11553)	2.32155	0.07802

In Table 3, based on JMLPC and CPHM we report the average of the predicted retention probabilities (ARP) across $\mathcal{T}$ at levels $j = 20$ and 30. The true retention proportions at level $j = 20$ and 30 in $\mathcal{T}$ are also reported. JMLPC produced 10.7% and 63.2% improvements respectively at $j = 20$ and 30 over CPHM. Two other discrepancy measures are also reported for evaluating the performance of CPHM and JMLPC on $\mathcal{T}$. We calculate the absolute deviations of the predicted probability $\hat{p}_{kj}$ that player k plays after level j from the true event $I\{L_k > j\}$ and report the mean absolute deviation across all players in $\mathcal{T}$: $\mathrm{MAD}(j) = \sum_{k\in\mathcal{T}}\hat{p}_{kj}\cdot \mathrm{I}(L_k \le j) + (1-\hat{p}_{kj})\cdot \mathrm{I}(L_k > j)$. We also reported the estimated odds ratio for true retention and dropouts as:

$$\mathrm{OR}(j) = \frac{\sum_{k\in\mathcal{T}}\hat{p}_{kj}/(1-\hat{p}_{kj})\cdot \mathrm{I}(L_k > j)}{\sum_{k\in\mathcal{T}}\hat{p}_{kj}/(1-\hat{p}_{kj})\cdot \mathrm{I}(L_k \le j)}\,.$$

Note that, while lower values of MAD signifies better predictive performance, higher values of OR is preferred. For $j = 20$ and 30, JMLPC respectively produces 30.4% and 62% improvement in OR over CPHM; the corresponding improvements in MAD is 7.1% and 41.5%. These shows that modeling the inter-dependencies among the player motivations in JMLPC is beneficial.

Next, we conduct five-step ahead predictions for engagements of players in $\mathcal{T}$ at levels $11,\ldots,15$ based on their history up to level 10. We report the root mean square error (RMSE) of the log-engagement predictions in table 4 along with the average log-playingtime (ALPT). Compared to empirical distribution based predictions, JMLPC produces 7.9% improvement in RMSE for one-step ahead prediction for level 11. The improvement is much more pronounced for multi-step ahead prediction; for level 15, we observe an improvement of 17.3%. In figure 10, based on gaming history till level 10, we report the dropout detection rates in JMLPC prediction in the future levels as well as the false positive rate in dropout prediction. The left panel of figure 10 shows that barring the next two levels the JMLPC fitted model provides good coverage in correctly predicting dropouts. In the right panel of figure 10, we report the false positive rate (FP) in dropout prediction at level l by the percentage of players predicted by our JMLPC model to have dropped out at least 3 levels before or 3 levels after than when they actually dropped out, i.e.,

$$\text{FP}_l = 1 - \frac{\text{number of players predicted by JMLPC to drop out in } [l-2, l+2]}{\text{number of actual dropouts in the test data at level } l}.$$

In figure 10 we observed that FP_l increases as l moves further away from level 10. This is expected as the predictive difficultly for $(l-10)$th step ahead prediction increases with l. We found that FP_l values were reasonably controlled for an appreciable range of future levels.

Table 4: One to five step ahead prediction of playing time (in log sec) from level 10 using the estimated JMLPC model

Level	11		12		13		14		15	
	ALPT	RMSE	ALPT	RMSE	ALPT	RMSE	ALPT	RMSE	ALPT	RMSE
Truth	8.0212	-	7.8224	-	7.6844	-	7.4354	-	7.8439	-
JMLPC	8.2865	1.0839	7.7928	1.0118	7.7653	0.9976	7.7111	1.1275	7.6967	1.0928
Empirical	8.5709	1.1770	8.3429	1.1382	8.1922	1.1029	8.2706	1.3670	8.6230	1.3212

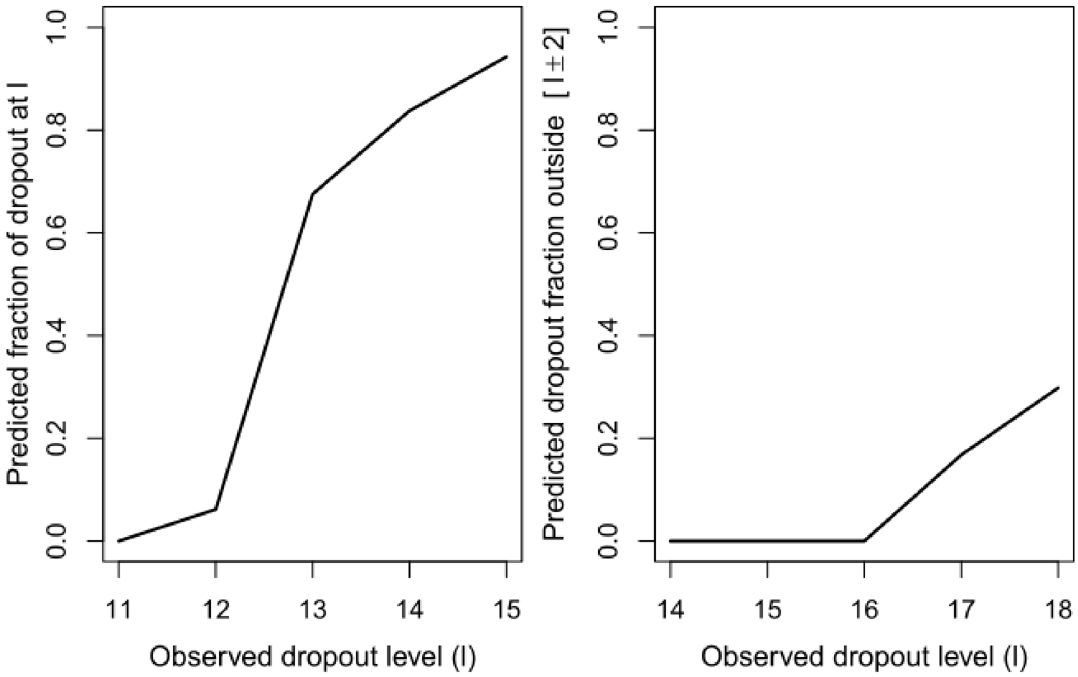


Figure 10: Out of sample dropout detection rate and false positive rate of JMLPC based on gaming history upto level 10.

The predicted lifetime engagement of a player k in $\mathcal{T}$ is given by:

$$\widehat{\mathrm{LE}}_k = \sum_{l=1}^{10} E_{kl} + \sum_{l=11}^{L} \hat{p}_{kl} \hat{E}_{kl} \ ,$$

where, E_{kl} is the known engagement duration of the player at level l; $\hat{E}_{kl}$ is a point-prediction of engagement duration and $\hat{p}_{kl}$ are the predicted retention probabilities. CPHM based LE predictions are computed using the predicted retention probability from CPHM model and engagement predictions based on the empirical distribution. The average of the logarithm of predicted LE values (ALLE) across all players in $\mathcal{T}$ as well as the RMSE of log LE predictions across different methods is reported in table 5. JMLPC yields 25.5% further reduction in RMSE over its nearest competitor.

Table 5: Prediction of lifetime player engagement based on playing history up to level 10

	ALLE	RMSE
Truth	10.53142	-
JMLPC	10.43483	0.39376
Cox Model	10.78096	0.52825
Empirical	10.68114	0.55780

6 Extending the JMLPC framework

The joint modeling approach JMLPC used in Section 4 and its associated estimation methodology are very flexible and can accommodate several variants and extensions of the model described in (1)-(6). First, to understand the role of the smoothness constraints on $\{\alpha_l^{(i)} : l = 1, \ldots, L\}$ and $\{\beta_{kl}^{(i)} : l = 1, \ldots, L\}$ consider the model JMLPC.UN which does not have any constraints on the $\alpha_l^{(i)}$s. The number of fixed effects from these unrestricted $\alpha_l^{(i)}$s in the joint model is $3L$ which equals 117 for the data set discussed in Section 2. While this large number of fixed effects increases computational time of the proposed algorithm, JMLPC.UN is still estimable in each split of the training data. Tables 6 and 7 show the performance of the model on the same test data as in Section 5.4.

Table 6: Performance of retention probability prediction by JMLPC.UN at future levels

Level	ARP	OR	MAD
20	0.07945	2.18342	0.12470
30	0.05391	2.92241	0.05391

Table 7: One step ahead prediction of playing time (in log sec) by variants of JMLPC

Level	Model	ALPT	RMSE
11	JMLPC.UN	8.01494	1.04189
	JMLPC.E	8.11175	1.15393

Comparing table 6 with table 3 and the first row of table 7 with table 4, we find that though JMLPC.UN provides a better prediction for playing time, its prediction for retention probabilities are worse than the constrained JMLPC model. Deterioration in predictive performance of the unconstrained model can be due to inefficient estimation of the model parameters. The smoothness structure across levels helps in reducing the intrinsic number of parameters and increases estimation accuracy by borrowing information across levels. Unconstrained models have lower relative signal strength due to larger number of free parameters which can lead to poor estimation of the model parameters. Note that, the random intercepts $\{\beta_{kl}^{(i)} : l = 1, \dots, L\}$ need to be constrained for identifiability. Incorporating more flexible structures than JMLPC on the random intercepts not only increases estimation complexity but may also cause convergence issues in the DR based estimation method. Non-convergence of the estimation algorithm can arise due to lack of data in the higher levels compared to the number of parameters which results in highly fluctuating coefficients estimates from the different splits.

Table 8: Estimated coefficients of the JMPLC.E model

log(engagement) (family = gaussian, link = identity)				
	Mean	St Dev	St Error	P-value
$a_1^{(1)}$	5.135	0.0885	0.0393	0
$a_2^{(1)}$	-1.1787	0.1102	0.0354	0
$a_3^{(1)}$	1.5157	0.1798	0.0701	0
$a_4^{(1)}$	-0.1657	0.0966	0.0354	0.022
gender: female	0.2024	0.0308	0.0093	0
afternoon	0.0162	0.0053	0.0006	0
evening	0.0341	0.0053	0.0006	0
city tier I	0.0871	0.074	0.0194	0.234
city tier II	-0.0068	0.0626	0.0247	0.992
city tier III	0.0254	0.0547	0.0151	0.66
city tier IV	-0.0911	0.063	0.0235	0.176
city tier V	0.1313	0.0615	0.0195	0.012
$\log(E_{\cdot,l-1})$	0.4528	0.0213	0.0094	0
$\log(A_{\cdot,l-1})$	-0.1681	0.0076	0.0012	0
$C_{\cdot,l-1}$	-0.0751	0.0369	0.0033	0.044

Collaboration (family = binomial, link = logit)				
	Mean	St Dev	St Error	P-value
$a_1^{(3)}$	-1.784	0.4083	0.2603	0
$a_2^{(3)}$	0.7983	0.4964	0.3025	0
$a_3^{(3)}$	0.3459	0.1252	0.0387	0
$a_4^{(3)}$	0.4995	0.1381	0.054	0
gender: female	1.7676	0.3098	0.1712	0
afternoon	0.0539	0.0557	0.0061	0.276
evening	0.0809	0.0551	0.0103	0.138
city tier I	-0.9624	0.3237	0.0764	0
city tier II	-2.5351	0.2322	0.0249	0
city tier III	-1.3504	0.2644	0.0481	0
city tier IV	-2.4431	0.2271	0.0392	0
city tier V	-1.6996	0.2787	0.0356	0
$\log(E_{\cdot,l-1})$	0.1301	0.0769	0.0078	0.09
$\log(A_{\cdot,l-1})$	-0.2943	0.0529	0.0042	0
$C_{\cdot,l-1}$	3.4701	0.2503	0.0773	0

log(Achievement) (family = gaussian, link = identity)				
	Mean	St Dev	St Error	P-value
$a_1^{(2)}$	8.1641	0.1225	0.0513	0
$a_2^{(2)}$	-0.4617	0.1111	0.0316	0
$a_3^{(2)}$	2.1299	0.178	0.0641	0
$a_4^{(2)}$	-0.6238	0.1117	0.0205	0
gender: female	-0.1834	0.0377	0.014	0
afternoon	-0.0003	0.0061	0.0004	0.968
evening	0.0352	0.0058	0.0006	0
city tier I	0.1802	0.0914	0.0197	0.038
city tier II	0.1639	0.0858	0.0307	0.006
city tier III	0.082	0.0771	0.0254	0.284
city tier IV	0.2808	0.0837	0.0304	0
city tier V	-0.0338	0.0824	0.0216	0.618
$\log(E_{\cdot,l-1})$	-0.1175	0.0205	0.0054	0
$\log(A_{\cdot,l-1})$	0.1828	0.0095	0.0014	0
$C_{\cdot,l-1}$	-0.3417	0.0504	0.0073	0

Survival Outcome:				
	Mean	St Dev	St Error	P-value
gender: female	0.4699	0.2153	0.0157	0.038
city tier I	0.3866	0.3014	0.0144	0.224
city tier II	-0.0611	0.1378	0.0046	0.67
city tier III	1.2126	0.1678	0.0058	0
city tier IV	0.5246	0.1358	0.0065	0
city tier V	0.5739	0.2037	0.013	0.002
Engagement: η_{11}	3.0826	2.9235	0.413	0.23
Engagement: η_{12}	0.4412	0.3532	0.0441	0.2
Engagement: η_{13}	1.0329	0.9161	0.1271	0.25
Engagement: η_{14}	-0.7546	0.7333	0.0802	0.286
Achievement: η_{21}	-1.6712	3.2562	0.3783	0.526
Achievement: η_{22}	-0.342	0.302	0.0261	0.254
Achievement: η_{23}	-0.5748	0.637	0.0807	0.384
Achievement: η_{24}	0.0872	0.34	0.0408	0.806
η_3	-0.3141	0.0367	0.0028	0

We next consider an extension JMLPC.E of JMLPC where we use all the available control variables such as gender, time of play, city tiers in (1)-(3). Additionally, we also use the player characteristics from the previous level $(E_{k,l-1}, A_{k,l-1}, C_{k,l-1})$ for predicting his/her responses at level l. JMLPC.E contains similar structural constraints on the intercepts as JMLPC. In table 8, we present the fitted JMLPC.E model. All the lagged variables were statistically significant. Most of the city tiers were not significant in predicting engagement or achievement but they were significant for predicting collaboration and survival outcomes.

A host of player specific gaming outcomes from the past levels such as $\{E_{k,j}, A_{k,j}, C_{k,j} : j = 1, \ldots, l-1\}$ can be used in the JMLPC framework. The usage of these lagged variables will however massively increase the computational costs associated with multi-step ahead predictions from the associated joint model. Also, using several such lagged variables will lead to multi-collinearity and reduced interpretability of the fitted model. Increasing model complexity by naively using several control variables will not necessarily lead to improved predictive performance as the variance of the model also increases. Comparing the RMSE in the last row of table 7 with the corresponding RMSE in table 4, we observe that JMLPC.E does worse in the prediction of playing time at the 11th level than JMLPC. Variable selection methods will be needed to properly select important variables and reduce variance in a joint modeling framework containing several controls and in-game lagged player attributes.

7 Conclusion

We develop a joint modeling framework to predict player retention as a player progresses through the different levels of the game. As we describe in the paper, level progression impacts player's motivation through enhanced gaming experience that occurs through collaboration and achievement. We capture players' motivations through level-wise engagement times, collaboration and achievement score and jointly model them using a generalized linear mixed model (glmm) framework that also encompasses a time-to-event variable corresponding to churn. By incorporating the level-wise changes in a player's motivations and using them for hazard rate prediction, our method greatly improves state-of-the-art retention models. In particular, we outperform aggregate statistics based methods that ignores level-wise progressions as well as progression tracking non-joint model such as the Cox proportional hazards method.

While our findings are based on data from a popular action based RPG, the model framework can be applied to increase customer retention in other freemium products that use inter-dependencies among users to increase users' stickiness (Appel et al., 2019, Ross, 2018). In particular, as our simulation results indicate that by offering different levels of promotion to enhance collaboration these freemium models can increase customer retention and revenue. Finally, other marketing retention metrics including player's lifetime gaming characteristics (Gupta et al., 2004) such as the lifetime engagement (cumulative playing time across levels) or collaborations, can be easily computed as functionals of these level-wise predictions. These metrics can offer more granular predictions of customer retention. This will help managers with decisions for offering more targeted player promotions to increase retention.

Table 9: Expected average retention probability (ERP) and expected playing time in seconds (EPT) of future players segmented by Gender and City-tiers across different levels of the game. The expected lifetime engagement (ELE) value is also reported in the last row of the table.

	Tier 1 City				Tier 2 City				Tier 3 City				Tier 4 City				Tier 5 City			
	Male		Female		Male		Female		Male		Female		Male		Female		Male		Female	
Level	ERP	EPT	ERP	EPT	ERP	EPT	ERP	EPT	ERP	EPT	ERP	EPT	ERP	EPT	ERP	EPT	ERP	EPT	ERP	EPT
1	1	270.06	1	348.43	1	248.56	1	346.72	1	255.94	1	364.49	1	251.77	1	349.89	1	262.13	1	353.67
2	0.9978	451.27	0.997	584.77	0.998	417.13	0.9976	579.74	0.9946	429.63	0.9941	606.93	0.9968	422.58	0.9963	585.56	0.9965	440.21	0.9958	590.91
3	0.994	741.85	0.992	965.19	0.9946	688.52	0.9935	953.63	0.9853	709.32	0.984	994.39	0.9912	697.58	0.99	964	0.9907	726.99	0.9886	971.32
4	0.9877	1179.5	0.9834	1541.3	0.9888	1099.3	0.9865	1518.2	0.9699	1132.8	0.9674	1577.4	0.9819	1113.9	0.9795	1535.6	0.9809	1161.2	0.9766	1545.1
5	0.9773	1786.9	0.9695	2340.4	0.9793	1669.9	0.9751	2300.1	0.945	1721.2	0.9408	2384	0.9667	1692.3	0.9624	2328.1	0.965	1764.4	0.9574	2339.8
6	0.9611	2533.8	0.9475	3325.9	0.9642	2373.8	0.9572	3264	0.9066	2447.5	0.9001	3377.8	0.9428	2405.9	0.9359	3305.4	0.9401	2508.1	0.9276	3319.9
7	0.9364	3310.9	0.9144	4349.9	0.941	3106.2	0.9301	4267.5	0.8506	3204.4	0.8418	4416.3	0.9068	3148.6	0.8964	4323.6	0.903	3281.1	0.8835	4341.8
8	0.9007	3921.7	0.8674	5152.2	0.9071	3680.9	0.8912	5058.7	0.7745	3802.3	0.764	5248.2	0.8557	3732.8	0.8411	5127.4	0.8505	3886.8	0.8225	5152.3
9	0.8521	4147.1	0.805	5439.5	0.8603	3887.5	0.8386	5352.5	0.6794	4027	0.6688	5583.5	0.7877	3946.2	0.769	5429.3	0.7812	4104.2	0.7438	5464.4
10	0.7894	3883.4	0.7273	5080.3	0.7991	3630.8	0.7713	5015.7	0.5699	3783.2	0.5615	5288.8	0.7031	3693.2	0.6813	5097.6	0.6954	3835.3	0.6493	5143.2
11	0.7028	3299.7	0.6248	4305.9	0.7131	3073.1	0.6789	4265.6	0.4397	3240.2	0.4365	4579.5	0.5919	3139.9	0.5684	4353.8	0.5835	3255.5	0.5305	4411.6
12	0.5839	2634.2	0.4921	3433.6	0.5922	2439.7	0.5531	3407.9	0.2948	2631.7	0.2994	3767.4	0.4497	2516	0.4277	3511.3	0.4414	2604.7	0.3873	3578.2
13	0.4531	2044.1	0.356	2669	0.4557	1881.6	0.4161	2647.2	0.174	2101.8	0.1845	3051.5	0.3083	1968.8	0.2912	2769.3	0.3014	2036	0.2542	2843.5
14	0.33	1594.2	0.2381	2091.3	0.3245	1457.4	0.2896	2067	0.0923	1706	0.1044	2510.2	0.1918	1556.2	0.1814	2208.6	0.1874	1608.4	0.1522	2289.6
15	0.2341	1282.2	0.1545	1695.8	0.2217	1165.2	0.1941	1666.5	0.0476	1437.8	0.058	2142.5	0.1144	1273.8	0.1093	1825.4	0.1122	1316.5	0.0886	1908.6
16	0.1724	1078.3	0.1056	1436.7	0.1565	973.72	0.1353	1403.2	0.027	1263	0.035	1902.4	0.0723	1089.1	0.07	1574.1	0.0715	1125.5	0.0553	1660.7
17	0.1359	939.83	0.0788	1261.4	0.119	844.42	0.1022	1224	0.0177	1142.1	0.0239	1735	0.0508	962.77	0.0498	1400.8	0.0506	995.12	0.0386	1489.2
18	0.1146	843.98	0.0641	1138.8	0.0977	755.07	0.0837	1100.1	0.0132	1056.5	0.0183	1616.3	0.0396	874.54	0.0391	1280	0.0397	904.02	0.03	1367.7
19	0.1019	778.46	0.0556	1054.9	0.0854	694.07	0.073	1015.4	0.0109	997.12	0.0152	1534.4	0.0334	813.93	0.0332	1196.5	0.0337	841.55	0.0253	1283.7
20	0.094	736.15	0.0505	1001.5	0.0778	655.04	0.0665	961.16	0.0095	959.84	0.0135	1483.6	0.0298	775.18	0.0297	1144.1	0.0302	801.66	0.0225	1230.7
21	0.0887	715.73	0.0471	974.9	0.0728	634.94	0.0622	934.18	0.0087	943.77	0.0123	1463.5	0.0274	756.55	0.0274	1118.6	0.0279	782.57	0.0207	1207.1
22	0.0847	714.19	0.0445	975.02	0.069	632.74	0.0589	932.29	0.0081	950.79	0.0115	1475.4	0.0257	757.65	0.0257	1122.1	0.0262	784.28	0.0193	1213.1
23	0.0812	732.47	0.0423	1001.9	0.0657	648.47	0.056	955.58	0.0075	981.72	0.0108	1524.6	0.0242	779.18	0.0242	1155.3	0.0247	807.17	0.0182	1249.8
24	0.078	769.39	0.0403	1054.4	0.0628	680.98	0.0535	1003.2	0.0071	1034.9	0.0101	1605.2	0.0229	819.86	0.0229	1215.4	0.0234	849.94	0.0171	1315.3
25	0.0752	825.23	0.0386	1131.1	0.0603	730.43	0.0513	1075.4	0.0067	1110.1	0.0096	1718.3	0.0217	879.56	0.0218	1302.9	0.0223	912.55	0.0162	1409.3
26	0.0727	903.17	0.037	1239.8	0.058	800.07	0.0492	1176.9	0.0063	1211.7	0.0091	1870.7	0.0207	962.04	0.0208	1423.9	0.0212	999.16	0.0155	1538.2
27	0.0705	1010.7	0.0356	1388.3	0.0559	896.49	0.0474	1316.5	0.006	1351	0.0087	2075.8	0.0198	1075.8	0.0199	1588.9	0.0203	1118.3	0.0147	1714.7
28	0.0684	1154.5	0.0344	1586.8	0.054	1025.3	0.0458	1502.2	0.0057	1534.3	0.0083	2344	0.019	1226.8	0.0191	1808.3	0.0195	1276.6	0.0141	1947
29	0.0666	1338.7	0.0332	1840.3	0.0523	1191	0.0443	1738.8	0.0055	1764.6	0.0079	2682	0.0182	1419	0.0184	2086.2	0.0188	1478.2	0.0135	2238.6
30	0.0649	1572.2	0.0322	2156.2	0.0507	1401.9	0.0429	2040.8	0.0053	2050.4	0.0076	3097.2	0.0176	1660.6	0.0177	2433.5	0.0181	1732.3	0.013	2602.8
31	0.0633	1869.2	0.0312	2567.6	0.0493	1670.4	0.0417	2422.7	0.0051	2405.8	0.0073	3608	0.017	1965.5	0.0171	2868.6	0.0175	2052.7	0.0125	3058.1
32	0.0619	2245.9	0.0303	3082.3	0.048	2013.2	0.0405	2907.8	0.0049	2848.9	0.0071	4238.5	0.0164	2350.1	0.0165	3416.8	0.0169	2457.5	0.0121	3625.4
33	0.0606	2727.2	0.0295	3738.8	0.0468	2451.7	0.0395	3526	0.0047	3404.1	0.0068	5022.1	0.0159	2837.7	0.016	4106.5	0.0164	2971.6	0.0117	4340.5
34	0.0594	3343.3	0.0287	4580.1	0.0456	3016	0.0384	4317	0.0045	4104.9	0.0066	6001.9	0.0154	3460.2	0.0155	4984.1	0.0159	3628.5	0.0113	5242.1
35	0.0582	4140.2	0.028	5664.6	0.0446	3745.7	0.0375	5336.3	0.0044	4995.3	0.0064	7237.8	0.0149	4258.4	0.0151	6103.3	0.0155	4472.1	0.011	6390.3
36	0.0572	5157.2	0.0274	7051.9	0.0436	4683.9	0.0366	6637.6	0.0042	6115.8	0.0062	8776.2	0.0145	5274.8	0.0146	7521.5	0.0151	5547.6	0.0107	7835.6
37	0.0562	6446.3	0.0267	8802.3	0.0426	5876.2	0.0358	8285.4	0.0041	7502.5	0.006	10664	0.0141	6552.1	0.0143	9292.7	0.0147	6900.5	0.0104	9634.1
38	0.0553	8069	0.0261	11001	0.0417	7379.7	0.035	10347	0.004	9203.4	0.0059	12955	0.0138	8142.1	0.0139	11487	0.0143	8589.6	0.0101	11842
39	0.0544	10095	0.0256	13800	0.0409	9303.2	0.0343	12987	0.0039	11275	0.0057	15741	0.0134	10147	0.0135	14242	0.014	10726	0.0098	14611
ELE	29432.66		34011.34		26775.34		35621.92		20072.24		28089.60		23237.95		31504.87		24034.94		30390.89	

Table 10: Expected retention probability (ERP) of future players and the expected playing times of those retained (EPT) based on mild, aggressive and no promotion campaigns for increasing in-game collaboration. The average premium purchase value (PP Value) (in cents) per one minutes playing duration in each level is also reported. The expected premium purchase (EPPV) value per future players is presented for each promotion type.

Level	**PP Value**	**None**			**Mild**			**High**		
No.	(cents/min)	**ERP**	**EPT**	**EPPV**	**ERP**	**EPT**	**EPPV**	**ERP**	**EPT**	**EPPV**
1	0.33992	1.00000	258.07	1.46	1.00000	257.99	1.46	1.00000	257.99	1.46
2	0.32978	0.99687	432.95	2.37	0.99735	432.82	2.37	0.99798	432.82	2.37
3	0.36541	0.99151	714.40	4.31	0.99282	714.19	4.32	0.99452	714.19	4.33
4	0.35798	0.98254	1140.35	6.68	0.98522	1140.01	6.70	0.98870	1140.01	6.72
5	0.29813	0.96792	1731.91	8.33	0.97282	1731.39	8.37	0.97918	1731.39	8.42
6	0.29377	0.94498	2461.65	11.39	0.95329	2460.91	11.49	0.96410	2460.91	11.62
7	0.31113	0.91059	3221.22	15.21	0.92382	3220.26	15.43	0.94116	3220.26	15.72
8	0.29123	0.86195	3818.60	15.98	0.88172	3817.47	16.34	0.90794	3817.47	16.82
9	0.29069	0.79762	4036.98	15.60	0.82526	4035.83	16.14	0.86256	4035.83	16.87
10	0.36985	0.71784	3778.46	16.72	0.75397	3777.45	17.56	0.80384	3777.45	18.72
11	0.35381	0.61307	3212.00	11.61	0.65811	3211.24	12.46	0.72230	3211.24	13.68
12	0.40232	0.47844	2571.94	8.25	0.53066	2571.44	9.15	0.60876	2571.44	10.50
13	0.37902	0.34211	2009.96	4.34	0.39565	2009.69	5.02	0.48079	2009.69	6.10
14	0.21507	0.22568	1585.41	1.28	0.27409	1585.32	1.56	0.35665	1585.32	2.03
15	0.23604	0.14402	1294.46	0.73	0.18383	1294.49	0.94	0.25667	1294.49	1.31
16	0.28500	0.09674	1104.15	0.51	0.12868	1104.27	0.67	0.19073	1104.27	1.00
17	0.49027	0.07123	974.33	0.57	0.09760	974.50	0.78	0.15123	974.50	1.20
18	0.53567	0.05734	883.86	0.45	0.08015	884.05	0.63	0.12804	884.05	1.01
19	0.45001	0.04946	821.75	0.30	0.07002	821.97	0.43	0.11419	821.97	0.70
20	0.46275	0.04471	782.17	0.27	0.06385	782.40	0.39	0.10559	782.40	0.64
21	0.24412	0.04163	762.98	0.13	0.05982	763.22	0.19	0.09990	763.22	0.31
22	0.20998	0.03929	763.89	0.11	0.05673	764.14	0.15	0.09551	764.14	0.26
23	0.30820	0.03727	785.42	0.15	0.05405	785.69	0.22	0.09167	785.69	0.37
24	0.33513	0.03549	826.21	0.16	0.05169	826.49	0.24	0.08828	826.49	0.41
25	0.29015	0.03393	886.24	0.15	0.04960	886.54	0.21	0.08524	886.54	0.37
26	0.75407	0.03253	969.36	0.40	0.04773	969.68	0.58	0.08252	969.68	1.01
27	0.67586	0.03128	1083.85	0.38	0.04605	1084.19	0.56	0.08005	1084.19	0.98
28	0.75244	0.03015	1235.89	0.47	0.04452	1236.26	0.69	0.07781	1236.26	1.21
29	0.67081	0.02913	1429.51	0.47	0.04314	1429.90	0.69	0.07576	1429.90	1.21
30	0.51098	0.02819	1673.27	0.40	0.04187	1673.68	0.60	0.07387	1673.68	1.05
31	0.63375	0.02734	1980.95	0.57	0.04070	1981.38	0.85	0.07213	1981.38	1.51
32	0.68962	0.02655	2369.55	0.72	0.03963	2369.99	1.08	0.07052	2369.99	1.92
33	0.95039	0.02582	2862.68	1.17	0.03863	2863.10	1.75	0.06902	2863.10	3.13
34	0.74168	0.02514	3492.38	1.09	0.03770	3492.75	1.63	0.06762	3492.75	2.92
35	0.57502	0.02450	4300.79	1.01	0.03683	4301.06	1.52	0.06630	4301.06	2.73
36	0.45089	0.02391	5331.14	0.96	0.03602	5331.27	1.44	0.06507	5331.27	2.61
37	0.34155	0.02336	6627.32	0.88	0.03525	6627.20	1.33	0.06390	6627.20	2.41
38	0.69716	0.02283	8243.97	2.19	0.03453	8243.41	3.31	0.06280	8243.41	6.02
39	0.73132	0.02234	9716.20	2.65	0.03385	9714.88	4.01	0.06176	9715.04	7.31

References

Appel, G., B. Libai, E. Muller, and R. Shachar (2019). On the monetization of mobile apps. *International Journal of Research in Marketing*.

Badrinarayanan, V. A., J. J. Sierra, and K. M. Martin (2015). A dual identification framework of online multiplayer video games: The case of massively multiplayer online role playing games (mmorpgs). *Journal of Business Research 68*(5), 1045–1052.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Banerjee, T., G. Mukherjee, S. Dutta, and P. Ghosh (2019). A large-scale constrained joint modeling approach for predicting user activity, engagement, and churn with application to freemium mobile games. *Journal of the American Statistical Association*, 1–29.

Battey, H., J. Fan, H. Liu, J. Lu, and Z. Zhu (2015). Distributed estimation and inference with statistical guarantees. *arXiv preprint arXiv:1509.05457*.

Borbora, Z., J. Srivastava, K.-W. Hsu, and D. Williams (2011). Churn prediction in mmorpgs using player motivation theories and an ensemble approach. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pp. 157–164. IEEE.

Bowman, S. L. (2010). *The functions of role-playing games: How participants create community, solve problems and explore identity*. McFarland.

Castro, E. G. and M. S. Tsuzuki (2015). Churn prediction in online games using players' login records: A frequency analysis approach. *IEEE Transactions on Computational Intelligence and AI in Games 7*(3), 255–265.

Chen, X. and M.-g. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 1655–1684.

Chris, C. (1982). A taxonomy of computer games. *The Art of Computer Game Design*. Available at https://www.statista.com/study/24719/mmo-gaming-statista-dossier/.

Clements, R. (2012). Rpgs took over every video game genre. Available at https://www.ign.com/articles/2012/12/12/rpgs-took-over-every-video-game-genre.

Coussement, K. and D. Van den Poel (2008). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications 34*(1), 313–327.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*(2), 187–202.

East, R., K. Hammond, and P. Gendall (2006). Fact and fallacy in retention marketing. *Journal of Marketing Management 22*(1-2), 5–23.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association 72*(359), 557–565.

Evans, E. (2016). The economics of free: Freemium games, branding and the impatience economy. *Convergence 22*(6), 563–580.

Fields, T. (2014). *Mobile & social game design: Monetization methods and mechanics*. CRC Press.

Gao, K. (2017). *Scalable Estimation and Inference for Massive Linear Mixed Models With Crossed Random Effects*. Ph. D. thesis, STANFORD UNIVERSITY.

Gao, K., A. Owen, et al. (2017). Efficient moment calculations for variance components in large unbalanced crossed random effects models. *Electronic Journal of Statistics 11*(1), 1235–1296.

Gupta, S., D. R. Lehmann, and J. A. Stuart (2004). Valuing customers. *Journal of marketing research 41*(1), 7–18.

Hill, S. (2019). Games rule the itunes app store: Most popular genres revealed. Available at https://www.gamasutra.com/.

Huang, Y., S. Jasin, and P. Manchanda (2019). "level up": Leveraging skill and engagement to maximize player game-play in online video games. *Information Systems Research 30*(3), 927–947.

Jerath, K., P. S. Fader, and B. G. Hardie (2011). New perspectives on customer "death" using a generalization of the pareto/nbd model. *Marketing Science 30*(5), 866–880.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

Jordan, M. I. (2012). Divide-and-conquer and statistical inference for big data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 4–4.

Jordan, M. I., J. D. Lee, and Y. Yang (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681.

Kawale, J., A. Pal, and J. Srivastava (2009). Churn prediction in mmorpgs: A social influence based approach. In *2009 International Conference on Computational Science and Engineering*, Volume 4, pp. 423–428. IEEE.

Kumar, V. (2014). Making" freemium" work. *Harvard business review 92*(5), 27–29.

Liu, C. Z., Y. A. Au, and H. S. Choi (2014). Effects of freemium strategy in the mobile app market: An empirical study of google play. *Journal of Management Information Systems 31*(3), 326–354.

Liu, P., T. Chan, and H. Che (2020). The pursuit of leadership in a multiplayer online role-playing game and its effect on player spending. available from https://www.scu.edu/business/marketing/faculty/pliu/.

Morik, K. and H. Köpcke (2004). Analysing customer churn in insurance data–a case study. In *European conference on principles of data mining and knowledge discovery*, pp. 325–336. Springer.

Mozer, M. C., R. Wolniewicz, D. B. Grimes, E. Johnson, and H. Kaushansky (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on neural networks 11*(3), 690–696.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo 2*(11), 2.

Niculescu, M. F. and D. J. Wu (2011). When should software firms commercialize new products via freemium business models. *Under Review*.

Nie, G., G. Wang, P. Zhang, Y. Tian, and Y. Shi (2009). Finding the hidden pattern of credit card holder's churn: A case of china. In *International Conference on Computational Science*, pp. 561–569. Springer.

Papaspiliopoulos, O., G. O. Roberts, and G. Zanella (2020). Scalable inference for crossed random effects models. *Biometrika 107*(1), 25–40.

Park, E., R. Rishika, R. Janakiraman, M. B. Houston, and B. Yoo (2018). Social dollars in online communities: The effect of product, user, and network characteristics. *Journal of Marketing 82*(1), 93–114.

Periáñez, Á., A. Saas, A. Guitart, and C. Magne (2016). Churn prediction in mobile social games: Towards a complete assessment using survival ensembles. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 564–573. IEEE.

Pierre-Louis, S. (2019). Essential facts about the computer and video game industry. Available at https://www.theesa.com/wp-content/uploads/2019/05/2019-Essential-Facts-About-the-Computer-and-Video-Game-Industry.pdf.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC press.

Rizopoulos, D. (2016). The R package JMbayes for fitting joint models for longitudinal and time-to-event data using mcmc. *Journal of Statistical Software 72*(7), 1–45.

Rizopoulos, D. and E. Lesaffre (2014). Introduction to the special issue on joint modelling techniques. *Statistical methods in medical research 23*(1), 3–10.

Rizopoulos, D., G. Verbeke, and E. Lesaffre (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 637–654.

Rizopoulos, D., G. Verbeke, and G. Molenberghs (2008). Shared parameter models under random effects misspecification. *Biometrika 95*(1), 63–74.

Rizopoulos, D., G. Verbeke, and G. Molenberghs (2010). Multiple-imputation-based residuals and diagnostic plots for joint models of longitudinal and survival outcomes. *Biometrics 66*(1), 20–29.

Rosenberg, L. J. and J. A. Czepiel (1984). A marketing approach for customer retention. *Journal of consumer marketing*.

Ross, N. (2018). Customer retention in freemium applications. *Journal of Marketing Analytics 6*(4), 127–137.

Statista (2018). Statista dossier on mmo and moba gaming. *Statista*. Available at https://www.statista.com/study/24719/mmo-gaming-statista-dossier/.

Vonesh, E. F., T. Greene, and M. D. Schluchter (2006). Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in medicine 25*(1), 143–163.

Wei, Y., W. Zhang, S. Yang, and X. Chen (2019). Online communities and social network structure. *Available at SSRN 3420525*.

Xie, Y., X. Li, E. Ngai, and W. Ying (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications 36*(3), 5445–5449.

Yang, Z. and R. T. Peterson (2004). Customer perceived value, satisfaction, and loyalty: The role of switching costs. *Psychology & Marketing 21*(10), 799–822.

Zhang, C., C. W. Phang, Q. Wu, and X. Luo (2017). Nonlinear effects of social connections and interactions on individual goal attainment and spending: Evidences from online gaming markets. *Journal of Marketing 81*(6), 132–155.

Zhang, X., Y. Zhou, Y. Ma, B.-C. Chen, L. Zhang, and D. Agarwal (2016). Glmix: Generalized linear mixed models for large-scale response prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 363–372.