A Large-scale Constrained Joint Modeling Approach For Predicting User Activity, Engagement And Churn With Application To Freemium Mobile Games

Trambak Banerjee[*], Gourab Mukherjee[*], Shantanu Dutta[†]and Pulak Ghosh[‡]

January 22, 2019[§]

Abstract

We develop a *Constrained Extremely Zero Inflated Joint* (CEZIJ) modeling framework for simultaneously analyzing player activity, engagement and drop-outs (churns) in app-based mobile freemium games. Our proposed framework addresses the complex interdependencies between a player's decision to use a freemium product, the extent of her direct and indirect engagement with the product and her decision to permanently drop its usage. CEZIJ extends the existing class of joint models for longitudinal and survival data in several ways. It not only accommodates extremely zero-inflated responses in a joint model setting but also incorporates domain-specific, convex structural constraints on the model parameters. Longitudinal data from app-based mobile games usually exhibit a large set of potential predictors and choosing the relevant set of predictors is highly desirable for various purposes including improved predictability. To achieve this goal, CEZIJ conducts simultaneous, coordinated selection of fixed and random effects in high-dimensional penalized generalized linear mixed models. For analyzing such large-scale datasets, variable selection and estimation is conducted via a distributed computing based split-and-conquer approach that massively increases scalability and provides better predictive performance over competing predictive methods. Our results reveal co-dependencies between varied player characteristics that promote player activity and engagement. Furthermore, the predicted churn probabilities exhibit idiosyncratic clusters of player profiles over time based on which marketers and game managers can segment the playing population for improved monetization of app-based freemium games.

Keywords: Constrained Joint Modeling; Mobile Apps; Freemium Behavior; Zero-inflation; Large-scale longitudinal data analysis; Dropouts; CEZIJ.

[§] The research here was partially supported by NSF DMS-1811866, Indian Institute of Management-Bangalore Challenge grant and the Zumberge individual award from the University of Southern California. Corresponding author: gmukherj@marshall.usc.edu

[*] Department of Data Sciences and Operations, University of Southern California, Los Angeles

[†] Department of Marketing, University of Southern California, Los Angeles

[‡] Department of Decision Sciences and Information Systems, Indian Institute of Management, Bangalore

1 Introduction

Mobile games have become an integral part of modern life (Koetsier, 2015). While their almost ubiquitous presence is increasingly reshaping the recreational, socialization, educational and learning media (Statista (2018), see Ch 1 and 3 of Hwong (2016), Garg and Telang (2012)), the monetization policies associated with these new mobile apps is rapidly revolutionizing the digital marketing and advertisement space in information systems (Appel et al., 2017, Liu et al., 2014). As such mobile games (as per industry standards formally defined as any app-based game played on an Internet enabled mobile device such as tablets, phones, etc) currently comprise 42% of the market share of global gaming products (McDonald, 2017) and more than eight hundred thousand mobile games were available for download in the iOS App Store alone, with approximately four hundred new submissions arriving each day (PocketGamer, 2018). To understand how quickly the gaming market is growing, a new industry study from Spil Games (Diele, 2013) reports that 1.2 billion people are now playing games worldwide, with 700 million of those online. The unprecedented growth and popularity of mobile games has resulted in a market with some very unique consumer characteristics (Boudreau et al., 2017). It is an extremely crowded market with significant proportion of revenue accumulated through advertisement based on free products (Appel et al., 2017). Specifically, app retention rates are much lower than the observed retention rates in classical products and services, with reports suggesting that more than 80% of all app users churn (drop out) within the first quarter (Perro, 2016, MarketingCharts, 2017). The freemium business model (Niculescu and Wu, 2011), which offers a certain level of service without cost and sells premium add-on components to generate revenue, is a popular strategy for monetization of these mobile games. As such, industry reports indicate that more than 90% of the mobile games start as free, and more than 90% of the profits currently come from products that began as free (AppBrain, 2017, Taube, 2013). User characteristics in freemium models differ in fundamental aspects from traditional marketing models. This necessitates development of new analytical methods for modeling freemium behavior.

1.1 Freemium model: Player Activity and Engagement

In the freemium market, firms initially attract customers with free usage of their products, with the expectation that free usage will lead customers to engage in future purchase of premium components. However, customers can always remain free users and never need to enjoy the premium components of the product. This is an important distinction with non-freemium business models, where customers *must* purchase in order to use the product. While the free to use part of freemium products helps to attract the consumer base quickly (Kumar, 2014), managers are uncertain on whether and how freemium can generate profits (Needleman and Loten, 2012) as majority of the consumers do not use the premium part of freemium. As such, unless a game is very popular, in-app purchases contribute an insignificant proportion of its revenue. Mobile marketing automation firm Swrve (Swrve, 2016) found that over 48% of all in-game revenue are derived from 0.19% of all players, which is a tiny segment. While in-game (direct) revenue is important, there are several indirect ways of monetizing the free users by involving them to engage with the game via social media (through facebook or twitter likes and posts of game achievements, inviting social media friends to join game, watching, liking or posting youtube videos related to the game) or the app center. To measure the daily engagement of a player, we judiciously combine her in-app purchases (direct source of revenue) with her varied involvements with the game in media (indirect source of monetization), under the notion that purchase is the highest form of engagement. We define a player's daily activity as the time she spends playing the game in the day. Positive daily activity does not always lead to positive engagement. It is commonly believed that as a game grows with increasing and prolonged player activities, it will have more positive as well as higher engagement values.

For game managers it is extremely important to accurately measure player activity, engagement and their co-dependencies. Also, varied retention strategies are often used to curb high churn rates and their effects need to be properly analyzed. Here, we develop a *Constrained Extremely Zero Inflated Joint* (CEZIJ) modeling framework that provides a disciplined statistical program for jointly modeling player activity, engagement and churn in online gaming platforms. Our proposed framework captures the co-dependencies between usage (activity), direct and indirect revenue (engagement), and dropouts (which is a time-to-event) and provides a systematic understanding of how the dependent variables influence each other and are influenced by the covariates. Furthermore, the CEZIJ framework can be used to predict the activity, engagement and attrition of new players. The ability to forecast behavior of new players is critical for managers, as this enables them to better predict the effectiveness of their gaming platform in engaging customers and thus attract future advertisers to their platform.

1.2 Joint modeling of player characteristics

Our joint modeling framework uses generalized linear mixed effect models (GLMM) and relies on a joint system of equations that model the relationships between activity, engagement, and churn. In the activity equations, we separately assess whether consumers are active (i.e. play the game) and the extent of their activity through the amount of time they spend playing the game. Engagement is modeled by the probability of having positive engagement and by a conditional model on the positive engagement values. In the churn equations, we account for permanent churn identified as those players who are not active for more than 30 consecutive days. Our modeling systems addresses the complex interdependencies between (1) the decision to use the free product, (2) how much time will be spent using the free product, (3) the decision to engage, (4) the extent of engagement and (5) the decision to churn. That is, the joint equation system comprehensively uncovers positive, negative, or zero co-dependencies among activity, engagement, and churn in freemium markets. In recent times, joint modeling of multiple outcomes have received considerable attention (Rizopoulos, 2012). Many applications consider the modeling of single or multiple longitudinal outcomes and a time-to-event outcome (e.g., Jiang (2007), McCulloch (2008), Rizopoulos et al. (2009, 2010), Banerjee et al. (2014), Rizopoulos and Lesaffre (2014)). Our motivation for jointly modeling the drivers of player gaming traits and dropout arises from the fact that there is heterogeneity across player's outcomes and one must combine these effects by correlating the multiple responses. Since these responses are measured on a variety of different scales (viz. time spend in hours, revenue in dollars), a flexible solution is to model the association between different responses by correlating the random heterogeneous effects from each of the responses. Such a model not only gives us a covariance structure to assess the strength of association between the responses, but also offers useful insights to managers, since despite huge popularity of mobile games among users, managers are not certain whether freemium is profitable. Furthermore, it is important for managers to understand how activity and engagement are related to player churn. While customers who frequently use the free product could be more satisfied, thus reducing their probability of churn (Gustafsson et al., 2005), free usage could be related to a greater probability of churn as there is little switching cost for customers due to their lower perceived value (Yang and Peterson, 2004). Earlier studies used simpler models for churn that are independent of the purchase rate (Jerath et al., 2011). Here we model churn allowing for possible co-dependencies with activity and engagement.

1.3 Statistical challenges

The online gaming data, which is the application case described in detail in section 2, and the particular business model of freemium, pose several statistical challenges and necessitates novel extensions of the joint modeling framework. We describe the details below.

(i) Extreme Zero-inflation - Freemium behavior suggests that even if a player is active on a day, it very rarely leads to purchases or social media engagement on her part. Thus, though both activity and engagement are zero-inflated, engagement has an extremely zero-inflated distribution. Mixture distributions of which zero-inflated distributions are a special case are commonly used in this kind of data. While there are multiple models that have been developed to accommodate data with excess zeros; see for example, Olsen and Schafer (2001), Min and Agresti (2005), Han and Kronmal (2006), Alfò et al. (2011), Greene (2009) and the references therein, there is not much attention on extreme zero-inflated data. Few recent works, e.g., Hatfield et al. (2012) show promise though. We develop a joint modeling framework that can accommodate extreme zero-inflation. The proposed framework allows us to accommodate large incidences of no-engagement by active players, such as that observed in freemium markets and helps managers more accurately forecast sales potential

for businesses with large active customer bases but small incidence of engagement by separating the confound between non-active and non-engaged. We highlight that this extreme zero inflated data is not only relevant to freemium markets but is also common in other businesses wherein a sizable portion of the active consumer base engages in very little purchase activity. For example, in the *online* setting, we may observe low incidences of online ratings (i.e. 1-5 star rating), user generated content creation, banner ad click-through, and search ad conversion (see for example Urban et al. (2013), Haans et al. (2013)). Likewise in the *offline* setting of purchase data for example, most product categories comprise less than 5% planned or actual purchase for an individual's visit to the grocery store (Hui et al., 2013). Thus, if managers are interested in assessing promotion on sales or individual level purchase activity in these contexts, we may be confronted with data that contains an extreme number of zeros.

(ii) Parametric Constraints - We develop a framework for incorporating domain specific structural constraints in our model for one may have prior knowledge that a vector of parameters lies on a simplex or follows a particular set of inequality constraints. It is quite common in gaming data to have prior information available on various activities of the player. For example, it is well-known that player characteristics will have a burgeoning weekend effect or marketers have prior knowledge on the comparative efficacies of the retention strategies particularly if they have known dosage demarcations. Using these side information is extremely important (James et al., 2013, Banerjee et al., 2018) and the CEZIJ framework incorporates these domain expertise though convexity constraints in our model.

(iii) Hierarchical Variable Selection - In online gaming data one usually encounters numerous covariates related to both game specific and player specific variables and choosing the relevant set of covariates is highly desirable for improving predictability. It is also important that the inferential problems associated with these data properly account for the presence of a lot of possibly spurious covariates. The high-dimensionality of these datasets, however, renders classical variable selection techniques incompetent. We develop a novel algorithm for estimation in the CEZIJ framework that conducts variable selection from a

large set of potential predictors in GLMM based joint model. To produce interpretable effects CEZIJ imposes a hierarchical structure on the selection mechanism and includes covariates either as fixed effects or composite effects where the latter are those covariates that have both fixed and random effects (Hui et al., 2017a)(See Section 4 for details). Efficient selection of fixed and random effect components in a mixed model framework has received considerable attention in recent years (Bondell et al. (2010), Fan and Li (2012), Lin et al. (2013); detailed background is provided in Section 4). Penalized quasi likelihood (PQL) approach has been used by Hui et al. (2017b) to conduct simultaneous (but non-hierarchical) selection of mixed effects in a GLMM framework with adaptive lasso and adaptive group lasso regularization. The CREPE (Composite Random Effects PEnalty) estimator of Hui et al. (2017a) conducts hierarchical variable selection in a GLMM with a single longitudinal outcome and employs a monte carlo EM (MCEM) algorithm of Wei and Tanner (1990) to maximize the likelihood. The CREPE estimator ensures that variables are included in the final model either as fixed effects only or as composite effects. Our proposed CEZIJ framework is related to Hui et al. (2017a) in its ability to conduct hierarchical variable selection in GLMMs. However, unlike Hui et al. (2017a), CEZIJ performs hierarchical variable selection in a joint model of multiple correlated longitudinal outcomes. Additionally, it can also incorporate any convexity constraint on the fixed effects.

(iv) Scalability - For any mobile game app, gargantuan volumes of user activity data are automatically accumulated. Analyzing such big datasets not only involves inferential problems associated with high-dimensional data analysis but also the computational challenges of processing large-scale (sample) longitudinal data. To process large longitudinal data-sets, CEZIJ leverages the benefits of distributed computing. Recently, algorithmic developments for increased scalability and reduced computational time without sacrificing the requisite level of statistical accuracy have received significant attention. See for example Jordan et al. (2013), Jordan et al. (2018), Lee et al. (2015) and the references therein. A popular approach is to conduct inference independently and simultaneously on K subsets of the full dataset and then form a global estimator by combining the inferential results from the K nodes in a computation-efficient manner. We take a similar approach for the

hierarchical selection of fixed and random effects by using the split-and-conquer approach of Chen and Xie (2014) that splits the original dataset into K non-overlapping groups, conducts variable selection separately in each group and uses a majority voting scheme in assimilating the results from the splits.

(v) Prediction and Segmentation - Predictive 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. Statistically, this necessitates construction of prognostic models that can not only forecast new user activity, engagement and drop-out behavior but also dynamically update such forecasts over time as new longitudinal information about them arrives. Based on our fitted joint model, we construct drop-out probability profiles (over time) for an out-of-sample generic player population and use them for segmentation of idiosyncratic player behaviors. Segmentation is a key analytical tool for managers. Users in different segments respond differently to varied marketing promotions. This enables managers to use relevant marketing promotions that better match user responses in different segments and increase efficiency of their marketing campaign.

We develop our joint modeling framework which accommodates all of the above mentioned extensions through an efficient and scalable estimation procedure. To the best of our knowledge, we are the first to study constrained joint modeling of high-dimensional data. Though we demonstrate the applicability of the CEZIJ inferential framework for the disciplined study of freemium behavior, it can be used in a wide range of other applications that needs analyzing multiple high-dimensional longitudinal outcomes along with a time-to-event analysis. To summarize, the key features of our CEZIJ framework are:

- Joint modeling of the highly related responses pertaining to daily player activity and engagement as well as the daily dropout probabilities using the freemium mobile game data described in Section 2;

- The possibility of acute zero-inflation in the player engagement distribution is addressed by modeling the conditional probability of no engagement given that the player had used the app in the day (see Figure 3);

- Convexity constraints pertinent to domain expertise and prior beliefs are incorporated in the modeling framework in Section 3;

- A penalized EM algorithm (Wei and Tanner, 1990) is used for simultaneous selection of fixed and random effects wherein data-driven weighted ℓ_1 penalties are imposed on the fixed effects as well as on the diagonal entries of the covariance matrix of the random effects while the common regularization parameter λ is chosen by a BIC-type criterion (see equation (9));

- Hierarchical selection of the fixed and random effects is conducted in Section 4 by using a re-weighted ℓ_1 minimization algorithm that alternates between estimating the parameters and redefining the data-driven weights such that the weights used in any iteration are computed from the solutions of the previous iteration;

- The divide and conquer approach in Section 5 distributes the problem into tractable parallel sub-groups resulting in increased scalability;

- Prediction of the drop-out probabilities as well as the activity and engagement characteristics of new players with the predictions being dynamically updated as additional longitudinal information is recorded (see Section 6). Based on these dynamic churn probability curves from our fitted joint model, we conduct segmentations of player profiles that can be used by game managers to develop improved promotion and retention policies specifically targeting different dominant player-types.

2 Motivating data: Activity, Engagement, Churn and Promotion Effects in Freemium Mobile Games

We consider daily player level gaming information for a mobile app game where users use robot avatars to fight other robots till one is destroyed. There were 38,860 players in our database and we tracked daily player level activity and purchases for 60 consecutive days starting from the release date of the game. We use a part of the data (players) for estimation and the other part as the hold out set for prediction (See details in Section 6).

There were three modes of the game and level progression can only be attained through the principal mode. However, the players get rewards (henceforth called in-game rewards) if they win games in all three modes. For the two non-principal modes, collecting rewards is the main objective. The players can use these rewards for improving their fighting equipments through upgrades of their existing inventories or in getting access to powerful new robots or for acquiring fancy game themes and background changes. The player can also buy these facilities (add-ons) using real money through direct in-app purchases (IAP). There were only 0.28% of the players who used real money for buying add-ons. The players are given premium rewards, which has much higher order of magnitudes than regular rewards, if they promote the game or the developers through social media (inviting friends on facebook for games, facebook likes, youtube likes, tweets) or through the app center or by downloading other related apps from the developer. Approximately 7.2% of the players in our data had premium rewards. We record daily engagement of a player by appropriately combining her real money purchases (direct source of revenue) with her varied involvement in promoting the game in media (indirect source of monetization) with the notion being that the highest form of engagement is the one leading to purchases. Daily engagement is an extremely zero-inflated variable. We assess player behavior in terms of her daily total playing time (activity), engagement value and drop out probability. We say that a player has dropped out if she has not logged-in for a month consecutively. For each player we have a host of time-dependent covariates generated through the game-play which we model as composite effects. They include current level of the game, number of games played daily in the three different modes of the game, how are the in-game rewards spent, etc (See table 3 and summary table 4 in section D of the supplementary material for details). From a gaming perspective, it is very interesting to study the effects on gaming time of the amount of in-game rewards that the players spend on either upgrading existing robots or purchasing new robots. Another interesting feature of the game, was the usage of "gacha" mechanism (Toto, 2012) which allowed the players to gamble in-game currency through lottery draws. The "gacha" is a very popular feature in freemium games (Kanerva, 2016). We use the currency employed by players in "gacha" as well as their gains, as covariates in

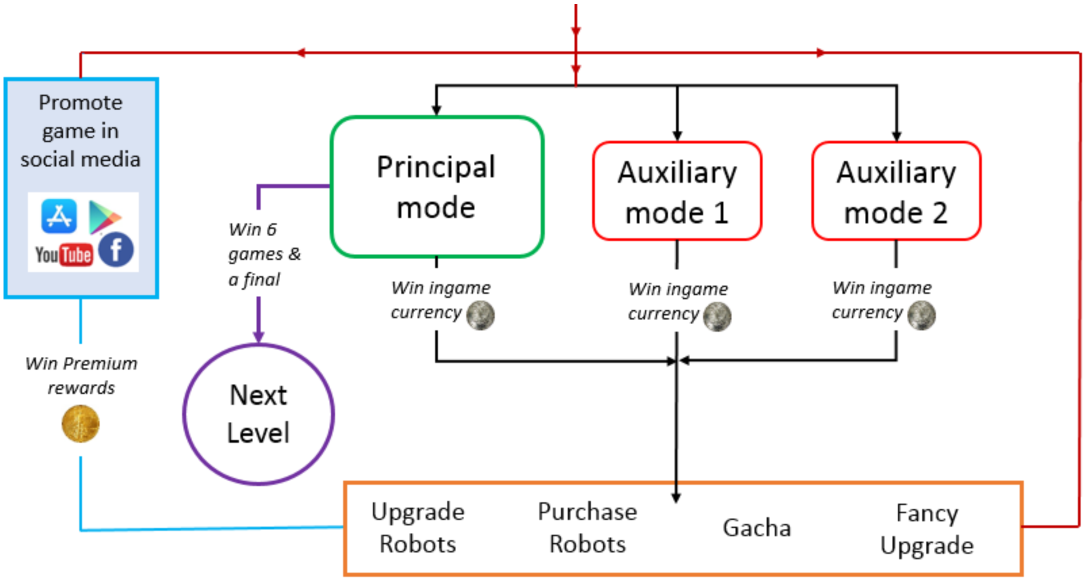
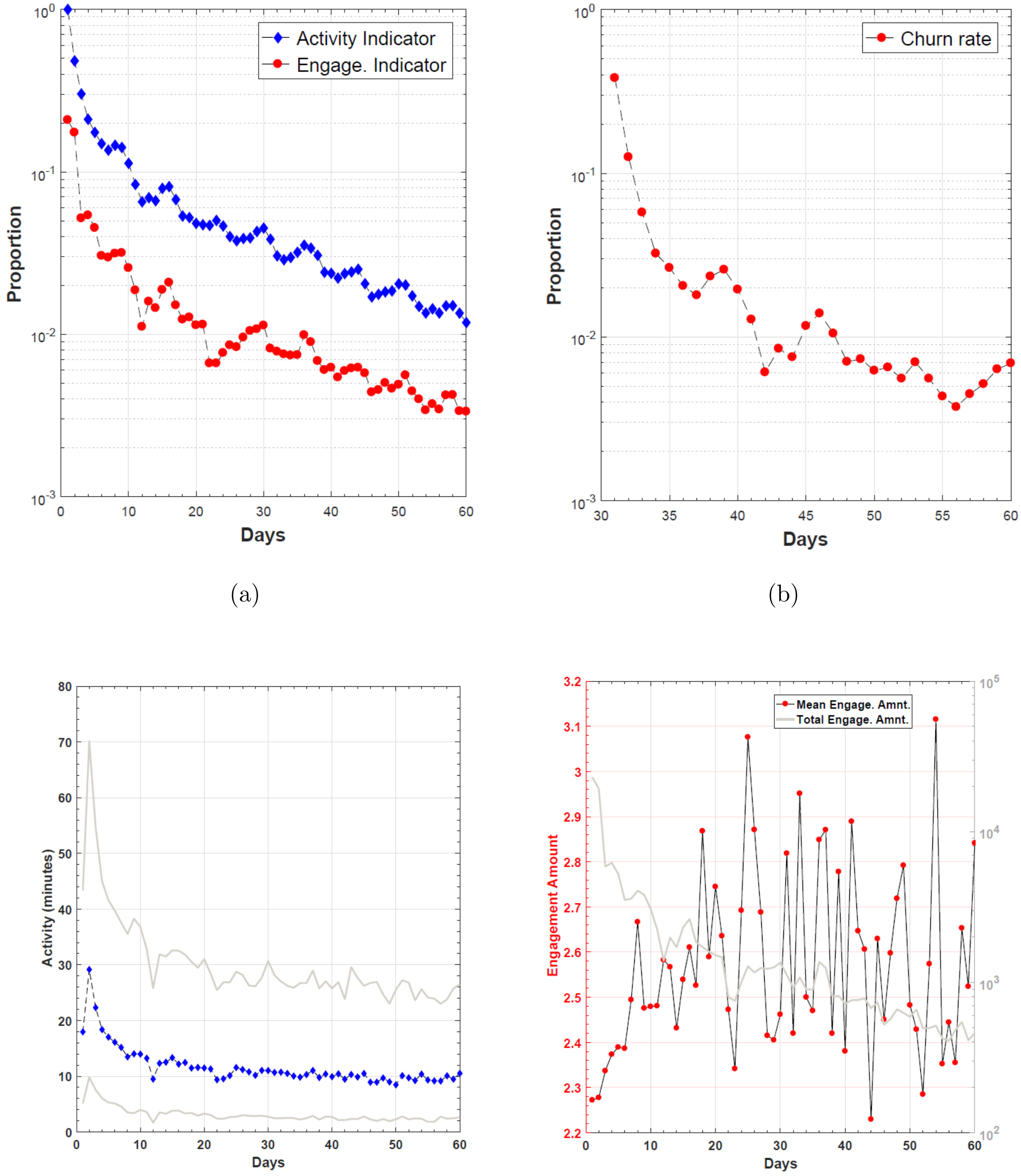


Figure 1: Game play flowchart

modeling engagement. Also, several promotional and retention strategies were used by the developers, which encourage player activity. Figure 1 contains a flowchart summarizing the key components of the game. The promotions intrinsically were of four different flavors: (a) award more reward percentages and battery life (b) sale on robots (c) thanksgiving holiday promotions (d) email and app-message based notifications for retention. Also, there were three different kinds of sales on robots. Thus, there were six different promotional strategies, with only one of them (if at all) being employed on a single day (See table 4 and figure 3 in section D of the supplementary material). In figures 2a and 2b, we present the activity, engagement and churn profiles of the players in our data. Interestingly, the proportion of players with positive engagement is below 10% from day 3 onwards and drops to less than 1% after the first 21 days. Figures 2c and 2d respectively show the 25^{th}, 50^{th} and 75^{th} percentiles of the distribution of Total Time Played and the average engagement amount on each of the 60 days. Note that from day 20 onwards the distribution of average engagement shows increased variability. This is not unexpected given the observation from figure 2a which shows that the proportion of players with positive engagement falls steadily. Also, note that the heavy tailed nature of the distributions of positive time played and positive engagement amount is evident from figure 2 (section D of the supplementary material)



(a)

(b)

(c)

(d)

Figure 2: (a) Proportion of players active and proportion of players with positive engagement over 60 days. (b) Proportion of player churn from day 31 to day 60. (c) Median activity sandwiched between its 25th and 75th percentile. (d) Mean engagement amount and the total engagement amount over the 60 days.

which plots the empirical CDF of the two variables. So, in the following section we use Log-normal distributions to model the non-zero activity and engagement values. Further details regarding the data are available in section D of the supplementary material.

3 CEZIJ Modeling Framework

Using the aforementioned motivation example, we now introduce our generic joint modeling framework. Consider data from n independent players where every player $i = 1, \ldots, n$ is observed over m time points. Let $\mathbb{A}_{ij}$ and $\mathbb{E}_{ij}$ denote, respectively, the activity and engagement of player i at day j with $\mathbb{A}_i = (\mathbb{A}_{i1}, \ldots, \mathbb{A}_{im})$ and $\mathbb{E}_i = (\mathbb{E}_{i1}, \ldots, \mathbb{E}_{im})$ denoting the corresponding vector of longitudinal measurements taken on player i. Let $\mathbb{D}_i$ denote the time of dropout for player i and $\mathbb{C}_i$ the censoring time. We assume $\mathbb{C}_i$s are independent of $\mathbb{D}_i$s. Thus $\mathbb{C}_i = m$ if player i never drops out. The observed time of dropout is $\mathbb{D}_i^* = \min(\mathbb{D}_i, \mathbb{C}_i)$, and the longitudinal measurements on any player i are available only over $m_i \leq \mathbb{D}_i^*$ time points. Suppose α_{ij} be the indicator of the event that player i is active ($\mathbb{A}_{ij} > 0$) on day j and ϵ_{ij} be the indicator that she positively engages ($\mathbb{E}_{ij} > 0$) on day j. Let $\pi_{ij} = \Pr(\alpha_{ij} = 1)$, $q_{ij} = \Pr(\epsilon_{ij} = 1 | \, \alpha_{ij} = 1)$, $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{im})$ and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{im})$. Note that, $\alpha_{ij} = 0$ implies $\mathbb{A}_{ij} = \mathbb{E}_{ij} = 0$ and also $\epsilon_{ij} = 0$. In these gaming apps, it is usually witnessed that any player's usage of the app always produces positive activity (however small). Thus, α_{ij} here corresponds to a player's daily activity indicator (AI). It forms the base (first level) of our joint model. The π_{ij} corresponds to daily usage probability where-as q_{ij} corresponds to the conditional probability of positive player engagement given that the player has used the app in the day. In Figure 3, we provide a schematic diagram of our joint model where we use two binary random variables: Activity Indicator (AI) and Engagement Indicator (EI) to be respectively denoted α_{ij} and ϵ_{ij}. We jointly model the five components $[\alpha_i, \mathbb{A}_i, \epsilon_i, \mathbb{E}_i, \mathbb{D}_i] := [\mathbb{Y}_i^{(s)} : s \in \{1, 2, 3, 4, 5\}]$ given the observations. Let $\mathcal{I}$ be the full set of p predictors in the data with $\mathcal{I}_f \subset \mathcal{I}$ as the set of fixed effects (time invariant or not) and $\mathcal{I}_c = \mathcal{I} \setminus \mathcal{I}_f$ as the set of composite effect predictors, which are modeled by combination of fixed and random effects. Let $p_f = |\mathcal{I}_f|$ and $p_c = |\mathcal{I}_c|$ and so, $p_c + p_f = p$. For each of the first four sub-models, $s = 1, \ldots, 4$, we consider p fixed effects $\boldsymbol{\beta}^{(s)}$ (p_f of those are from the time invariant and the rest from the composite components) and p_c random effects $\boldsymbol{b}^{(s)}$ while for the dropout model, $s = 5$, we consider p new fixed effects $\boldsymbol{\beta}^{(5)}$ but share the random effects from the four sub-models and calibrate their effects on dropouts through an association parameter vector $\boldsymbol{\eta}$. See section 3.1 for further details.

Let $x_{ijk}^{(s)}$ denote the observed kth covariate value for the ith player on the jth day. Let $\boldsymbol{x}_{ij}^{(s)} = \{x_{ijk}^{(s)} \mid k \in \mathcal{I}\}$ and $\boldsymbol{z}_{ij}^{(s)} = \{z_{ijk}^{(s)} \mid k \in \mathcal{I}_c\}$ denote the set of covariate values pertaining to the in-model fixed and random effects; $\mathbb{X}^{(s)}$ and $\mathbb{Z}^{(s)}$ respectively denote the data for these effects across all n players and $\boldsymbol{\beta} = \{\boldsymbol{\beta}^{(s)} : s \in \{1, 2, 3, 4\}\}$ and $\mathbb{b} = \{\boldsymbol{b}^{(s)} : s \in \{1, 2, 3, 4\}\}$ be all the fixed and random effects across all players. To join the four models, we take a correlated random effects approach and assume that the random effects governing the four sub-models have a multivariate Gaussian distribution. For player i, represent all her random effects by $\boldsymbol{b}_i = (\boldsymbol{b}_i^{(s)} : 1 \leq s \leq 4)$. We assume that $\{\boldsymbol{b}_i : 1 \leq i \leq n\}$ i.i.d. $N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is the $4p_c \times 4p_c$ unknown covariance matrix. To model the dropouts, we again consider p new fixed effects $\boldsymbol{\beta}^{(5)}$ but share the random effects from the four sub-models and calibrate their effects on dropouts through an association parameter vector $\boldsymbol{\eta}$. We model $\left[\mathbb{Y}^{(s)} : 1 \leq s \leq 5 \middle| \mathbb{X}, \mathbb{Z}, \boldsymbol{\beta}, \boldsymbol{\Sigma}\right]$ as

$$\prod_{i=1}^{n} \left[\, \boldsymbol{b}_i \,\right] \left[\alpha_i \,|\, \mathbb{X}_i^{(1)}, \boldsymbol{\beta}^{(1)}, \mathbb{Z}_i^{(1)}, \boldsymbol{b}_i^{(1)}\right] \left[\mathbb{A}_i \,|\, \alpha_i, \mathbb{X}_i^{(2)}, \boldsymbol{\beta}^{(2)}, \mathbb{Z}_i^{(2)}, \boldsymbol{b}_i^{(2)}\right] \left[\epsilon_i \,|\, \alpha_i, \mathbb{X}_i^{(3)}, \boldsymbol{\beta}^{(3)}, \mathbb{Z}_i^{(3)}, \boldsymbol{b}_i^{(3)}\right]$$
$$\left[\mathbb{E}_i \,|\, \alpha_i, \epsilon_i, \mathbb{X}_i^{(4)}, \boldsymbol{\beta}^{(4)}, \mathbb{Z}_i^{(4)}, \boldsymbol{b}_i^{(4)}\right] \left[\mathbb{D}_i \,|\, \mathbb{X}_i^{(5)}, \boldsymbol{\beta}^{(5)}, \boldsymbol{b}_i\right] .$$

Note that the dimension of each $\boldsymbol{b}_i^{(s)}$ in $\boldsymbol{b}_i$ is p_c and that of $\boldsymbol{x}_{ij}^{(s)}$ is p. In the context of our mobile app game data, $p_c = 25$ and so $\boldsymbol{\Sigma}$ is 100×100 and $p = 31$ for each of the five sub-models, thus making a set of 155 fixed effects (time invariant or not). See section 6 for more details.

Remark 1 If we have data pertaining to social media interactions among players, it would be beneficial to include network or group effects among players. In the absence of such network information, we model $\boldsymbol{b}_i^{(s)}$ as i.i.d. across players.

3.1 Longitudinal sub-models and model for Dropouts

Zero inflated Log-normal for modeling Activity - Since player i is active only at some time points j, the observed activity $\mathbb{A}_i$ has a mix of many zeros and positive observations. In equation (1), we consider a zero inflated (ZI) Log Normal model for $\mathbb{A}_{ij}$ to capture both the prevalence of these excess zeros and possible large values observed in the support of $\mathbb{A}_{ij}$. Thus, the model for activity $\mathbb{A}_{ij}$ has a mixture distribution with pdf

$$g_1(\alpha_{ij}, \mathbb{A}_{ij} \,|\, \boldsymbol{b}_i^{(1)}, \boldsymbol{b}_i^{(2)}) = (1-\pi_{ij})\,\mathbb{I}\{\alpha_{ij}=0\} + \pi_{ij}(\sigma_1 \mathbb{A}_{ij})^{-1}\phi\left(\frac{\log \mathbb{A}_{ij} - \mu_{ij}}{\sigma_1}\right)\mathbb{I}\{\alpha_{ij}=1\} \quad (1)$$

where

$$\begin{aligned} \mathsf{logit}(\pi_{ij}) &= \boldsymbol{x}_{ij}^{(1)T}\boldsymbol{\beta}^{(1)} + \boldsymbol{z}_{ij}^{(1)T}\boldsymbol{b}_i^{(1)} \quad \rightarrow \textbf{Binary part} && (2)\\ \text{and } \mu_{ij} &= \boldsymbol{x}_{ij}^{(2)T}\boldsymbol{\beta}^{(2)} + \boldsymbol{z}_{ij}^{(2)T}\boldsymbol{b}_i^{(2)} \quad \rightarrow \textbf{Positive activity part} && (3) \end{aligned}$$

The activity indicator α_{ij} is modeled using a logistic regression model with random effects in equation (2). In equation (3) we use an identity link to connect the expected log activity with the covariates and the random effects. For convenience, hereon the dependence on the fixed effects and covariates are kept implicit in the notations and only the involved random effects are explicitly demonstrated.

Extreme ZI Log-normal for modeling Engagement - Note that, $\mathbb{E}_i$ also has a mix of zeros and positive observations but the extreme zero events in the engagement variable are due to : (a) players are inactive on days and, (b) active players on a day may not exhibit engagement on the same day. To account for this excess prevalence of zeros, we use an Extreme Zero Inflated (EZI) Log Normal model that models $(\alpha_{ij}, \epsilon_{ij}, \mathbb{E}_{ij},)$ as a flexible mixture distribution with joint pdf

$$\begin{aligned} g_2(\alpha_{ij}, \epsilon_{ij}, \mathbb{E}_{ij} \,|\, \boldsymbol{b}_i^{(1)}, \boldsymbol{b}_i^{(3)}, \boldsymbol{b}_i^{(4)}) &= (1-\pi_{ij})\mathbb{I}\{\alpha_{ij}=0\} + \pi_{ij} g_3(\epsilon_{ij}, \mathbb{E}_{ij} \,|\, \boldsymbol{b}_i^{(3)}, \boldsymbol{b}_i^{(4)})\mathbb{I}\{\alpha_{ij}=1\} && (4)\\ \text{where, } g_3(\epsilon_{ij}, \mathbb{E}_{ij} \,|\, \boldsymbol{b}_i^{(3)}, \boldsymbol{b}_i^{(4)}) &= (1-q_{ij})\mathbb{I}\{\epsilon_{ij}=0\} + && (5)\\ & \quad q_{ij}(\sigma_2 \mathbb{E}_{ij})^{-1}\phi\left(\frac{\log \mathbb{E}_{ij} - \gamma_{ij}}{\sigma_2}\right)\mathbb{I}\{\epsilon_{ij}=1\} \\ \mathsf{logit}(q_{ij}) &= \boldsymbol{x}_{ij}^{(3)T}\boldsymbol{\beta}^{(3)} + \boldsymbol{z}_{ij}^{(3)T}\boldsymbol{b}_i^{(3)} \;\rightarrow \textbf{Binary part} && (6)\\ \gamma_{ij} &= \boldsymbol{x}_{ij}^{(4)T}\boldsymbol{\beta}^{(4)} + \boldsymbol{z}_{ij}^{(4)T}\boldsymbol{b}_i^{(4)} \rightarrow \textbf{Positive engagement part} && (7) \end{aligned}$$

Note that a player can potentially engage ($\mathbb{E}_{ij} \geq 0$) only if she is active ($\alpha_{ij} = 1$) on that day. Thus, $g_3(\epsilon_{ij}, \mathbb{E}_{ij} \,|\, \boldsymbol{b}_i^{(3)}, \boldsymbol{b}_i^{(4)})$ in equation (4) represents the joint pdf of $(\epsilon_{ij}, \mathbb{E}_{ij})$ conditional on the event that the player is active, i.e., $\alpha_{ij} = 1$. However, even if the player is active, distribution of engagement again can have a mixture distribution, as the particular player may or may not exhibit positive engagement ($\mathbb{E}_{ij} > 0$). Thus, conditional on the player being active, we further model $(\epsilon_{ij}, \mathbb{E}_{ij})$ using another zero-inflated Log Normal model as shown in equation (5). By combining equations (4) and (5), intuitively, we use the EZI model to split the players into two groups: (1) who are not active and (2) who are active. Then conditional on being active, we further split the latter group of players into two additional segments: (1) who do not engage ($\epsilon_{ij} = 0$) and, (2) who engage ($\epsilon_{ij} = 1$) and thus demonstrate positive engagement ($\mathbb{E}_{ij} > 0$). Finally, we complete the specification of the EZI Log Normal model by connecting the binary response $\epsilon_{ij}|\alpha_{ij} = 1$ with the covariates and the random effects through a logit link in equation (6) and use an identity link for expected log engagement γ_{ij} in equation (7). Note that even though we model α_{ij} in equations (1) and (4) using g_1 and g_2, respectively, there is no discordance as $g_1(\alpha_{ij}) = g_2(\alpha_{ij})$ for all (i, j).

Model for dropouts - For the discrete time hazard of dropout, we model $\lambda_{ij} := P(\mathbb{D}_i = j|\mathbb{D}_i \geq j,\, \boldsymbol{b}_i)$ as

$$\text{logit}(\lambda_{ij}) = \boldsymbol{x}_{ij}^{(5)T}\boldsymbol{\beta}^{(5)} + \boldsymbol{\eta}^T\boldsymbol{b}_i \ , \tag{8}$$

and the pmf of $\mathbb{D}_i^*$ is

$$g_4(\mathbb{D}_i^* = d \mid \boldsymbol{b}_i) = \left\{\prod_{j=1}^{d-1}(1-\lambda_{ij})\right\}\lambda_{id}^{\delta_i^{\mathbb{D}}}(1-\lambda_{id})^{1-\delta_i^{\mathbb{D}}}$$

where $\delta_i^{\mathbb{D}} = I(\mathbb{D}_i \leq \mathbb{C}_i)$ is the indicator of dropout occurrence. Here $\boldsymbol{\eta}$ is a parameter vector that relates the longitudinal outcomes and the dropout time via the random effects $\boldsymbol{b}_i$. This approach to modeling the dropouts through equation (8) is analogous to the shared parameter models in clinical trials that are used to account for potential Not Missing At Random (NMAR) responses. (see Vonesh et al. (2006), Guo and Carlin (2004) for example). If $\boldsymbol{\eta} = \boldsymbol{0}$ then the dropout is ignorable given the observed data. Figure 3 contains a schematic diagram of our joint model.

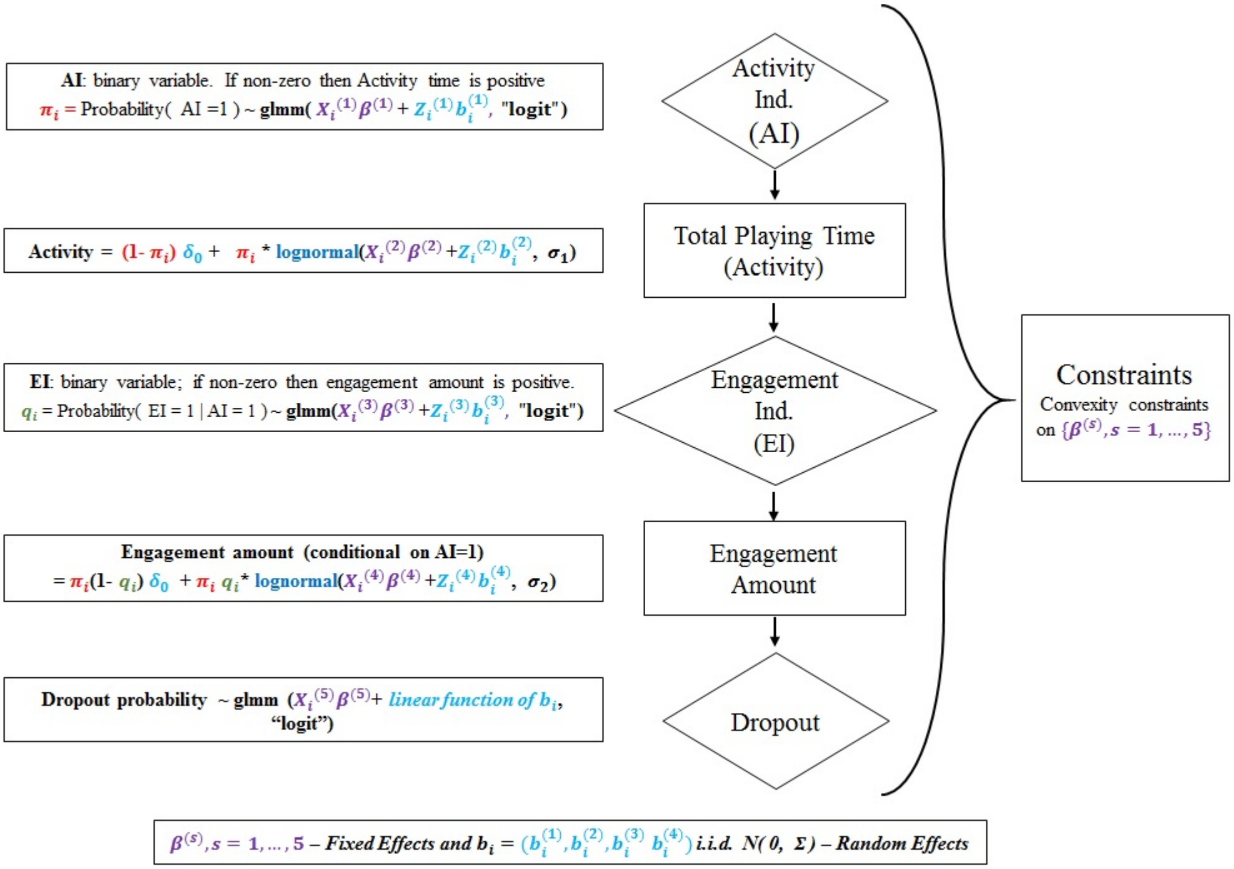

AI: binary variable. If non-zero then Activity time is positive
π_i = Probability(AI =1) ~ glmm($X_i^{(1)}\beta^{(1)} + Z_i^{(1)}b_i^{(1)}$, "logit")
Activity Ind. (AI)
Activity = (1- π_i) δ_0 + π_i * lognormal($X_i^{(2)}\beta^{(2)}+Z_i^{(2)}b_i^{(2)}$, σ_1)
Total Playing Time (Activity)
EI: binary variable; if non-zero then engagement amount is positive.
q_i = Probability(EI = 1 | AI = 1) ~ glmm($X_i^{(3)}\beta^{(3)}+Z_i^{(3)}b_i^{(3)}$, "logit")
Engagement Ind. (EI)
Constraints
Convexity constraints on $\{\beta^{(s)}, s = 1, \dots, 5\}$
Engagement amount (conditional on AI=1)
= π_i(1- q_i) δ_0 + π_i q_i* lognormal($X_i^{(4)}\beta^{(4)}+Z_i^{(4)}b_i^{(4)}$, σ_2)
Engagement Amount
Dropout probability ~ glmm ($X_i^{(5)}\beta^{(5)}$+ *linear function of* b_i, "logit")
Dropout
$\beta^{(s)}, s = 1, \dots, 5$ – *Fixed Effects and* $b_i = (b_i^{(1)}, b_i^{(2)}, b_i^{(3)}\ b_i^{(4)})$ *i.i.d.* $N(0,\ \Sigma)$ – *Random Effects*

Figure 3: Schematic diagram of our joint model for player i. The suffix denoting the day number is dropped for presentational ease.

Correlating the random effects and Linking the sub-models - All the sub-models described above carry information about the playing behavior of individuals and are therefore inter-related. To get the complete picture and to account for the heterogeneity across individual's outcomes, one must combine these effects by correlating the multiple outcomes. Without inter-relating or jointly considering these outcomes, it is not only hard to answer questions about how the evolution of one response (e.g., activity) is related to the evolution of another (e.g., engagement) or who is likely to dropout, but also problematic to model the heterogeneity. In such cases, it is natural to consider models where the dependency among the responses may be incorporated via the presence of one or more latent variables. A flexible solution is to model the association between different responses by correlating the random heterogeneous effects from each of the responses. In our joint modeling approach, random effects are assumed for each longitudinal response and they are associated by imposing a joint multivariate distribution on the random effects, i.e, $\boldsymbol{b}_i = (\boldsymbol{b}_i^{(s)} \ : \ 1 \leq s \leq 4) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Such a model borrows information across the various touch points and offers an intuitive way of describing the dependency between the responses. For example, questions such as, "is engagement related to activity for an individual?", or "does higher activity increase the probability of engagement" can be answered using the estimated covariance structure of $\boldsymbol{\Sigma}$. Furthermore, we assume that the dependency between the longitudinal outcomes and the risk of dropout are described by the random effects $\boldsymbol{b}_i$ and the covariates. In our context this is reasonable since, for instance, the longitudinal outcome AI may characterize player engagement, and player engagement can in turn influence the risk of dropout.

3.2 Parametric Constraints

The CEZIJ framework can incorporate any convexity constraints on the fixed effects: $\mathfrak{f}^{(s)}(\boldsymbol{\beta}^{(s)}) \leq 0$, $s = 1, \cdots, 5$, where $\mathfrak{f}$ is any pre-specified convex function. In the mobile game platform modeling application, domain expertise can be incorporated into our framework via these constraints. For example, industry belief dictates that all other factors remaining fixed, players have higher chance of being active in the game on weekends than on week days. Thus a sign constraint on the unknown fixed effect coefficient for the variable ($\beta^{(s)}_{\mathsf{weekend}} > 0$) that indicates whether day j is a weekend, is a simple yet effective way to include this additional information into our estimation framework. Also, different promotional and retention strategies used in these games are incorporated in the model as fixed effects through the binary variables demarcating the days they were applied (see figure 3 in section D of the supplementary material for a distribution of the various promotion strategies across the $m = 60$ days). These strategies often have previously known efficacy levels which imply monotonicity constraints on their effects. For example, email and app messaging based retention scheme should have at least a non-negative increment effect on the daily usage probabilities π_{ij}s; the engagement effect of a promotion that offers sale on only selected robots can not exceed the increment effect of sale on all robots. As such,

Table 1: Parameter constraints and their interpretation. Here $\beta^{(s)}_{prom(i)}$ indicates the fixed effect coefficient for promotion $i = I, \ldots, VI$ under model $s = 1, \ldots, 5$.

Constraint	Description
$\beta^{(s)}_{\mathsf{weekend}} \geq 0, \forall\ s$	Expect increased player activity on weekends
$\beta^{(s)}_{\mathsf{timesince}} \leq 0, \forall\ s$	Expect lower player activity as time since last login increases
$\beta^{(1)}_{\mathsf{promV}}, \beta^{(1)}_{\mathsf{promIV}} \geq 0$	Expect promotions IV, V to increase player activity
$\beta^{(1)}_{\mathsf{promV}} \geq \beta^{(1)}_{\mathsf{promIV}}$	Expect promotion V to have a higher positive impact on player activity than promotion IV
$\beta^{(2)}_{prom(i)} \geq 0$ for $i \neq IV$	All promotions other than IV to have a non-negative impact on activity.
$\beta^{(2)}_{promIII} \geq \beta^{(2)}_{\mathsf{promV}}$	Promotions III leads to a higher increase in activity than promotion V
$\beta^{(2)}_{\mathsf{promVI}} \geq \beta^{(2)}_{\mathsf{promII}} \geq \beta^{(2)}_{\mathsf{promV}}$	Promotions VI has the largest positive impact on activity followed by promotions II and V
$\beta^{(2)}_{\mathsf{promV}} \geq \beta^{(2)}_{promIV}$	Promotion V leads to a higher increase in activity than promotion IV

in our mobile game application, we assimilate these side information through structured affine constraints: $\mathbf{C}^{(s)}\boldsymbol{\beta}^{(s)} \leq \kappa^{(s)}$, $s = 1, \cdots, 5$ where $\mathbf{C}^{(s)}$ and $\kappa^{(s)}$ are known. Details about these constraints are provided in tables 1 and 5 (in section D of the supplementary material), where we describe the six promotion strategies and the constraints that have been included in our estimation framework along with their business interpretation.

4 Variable selection in CEZIJ

In the absence of any prior knowledge regarding variables that may appear in the true model, we conduct automated variable selection. Selection of fixed and random effect components in a mixed model framework has received considerable attention. Under the special case of a linear mixed effect model, Bondell et al. (2010) and Ibrahim et al. (2011) proposed penalized likelihood procedures to simultaneously select fixed and random effect components, while Fan and Li (2012), Peng and Lu (2012) and Lin et al. (2013) conduct selection of fixed and random effects using a two stage approach. Procedures to select only the fixed effects or the random effects have also been proposed under a GLMM framework; see Pan and Huang (2014) and the references therein. Simultaneous selection of fixed and random effect components in a GLMM framework is, however, computationally challenging.

The high dimensional integral with respect to the random effects in the marginal likelihood of GLMM often has no analytical form and several approaches have been proposed to tackle this computational hurdle: for example Laplacian approximations (Tierney and Kadane, 1986), adaptive quadrature approximations (Rabe-Hesketh et al., 2002), penalized quasi likelihood (PQL)(Breslow and Clayton, 1993) and EM algorithm (McCulloch, 1997). We use a penalized EM algorithm and for proper interpretation of composite effects we conduct joint variable selection of fixed and random effects in a hierarchical manner, which ensures that non-zero random effects in the model are accompanied by their corresponding non-zero fixed effects. Let $\boldsymbol{\Theta} = \big\{\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(5)}, \sigma_1, \sigma_2, \boldsymbol{\eta}, vec(\boldsymbol{\Sigma})\big\} := \big\{\boldsymbol{\theta}, vec(\boldsymbol{\Sigma})\big\}$ denote the vector of all parameters to be estimated. The marginal log-likelihood of the observed data under the joint model is:

$$\ell(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \int p\big(\alpha_i, \mathbb{A}_i, \epsilon_i, \mathbb{E}_i, \mathbb{D}_i^* |\, \boldsymbol{b}_i, \boldsymbol{\theta}\big)\, p\big(\boldsymbol{b}_i|\boldsymbol{\Sigma}\big)\, d\boldsymbol{b}_i = \sum_{i=1}^{n} \ell_i(\boldsymbol{\Theta}), \text{ where,}$$

$$\ell_i(\boldsymbol{\Theta}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| + \log \int \exp\Big(\sum_{j=1}^{m_i} \log p(\alpha_{ij}, \mathbb{A}_{ij}, \epsilon_{ij}, \mathbb{E}_{ij}, \mathbb{D}_i^* |\, \boldsymbol{b}_i, \boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{b}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{b}_i\Big) d\boldsymbol{b}_i$$

We estimate $\boldsymbol{\Theta}$ using the EM algorithm for Joint models (Rizopoulos, 2012) where we treat the random effects $\boldsymbol{b}_i$ as 'missing data' and obtain $\widehat{\boldsymbol{\Theta}}$ by maximizing the expected value of the complete data likelihood $\ell^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b})$ where

$$\ell^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}) = -\frac{n}{2}\log\boldsymbol{\Sigma} + \sum_{i=1}^{n}\Big(\sum_{j=1}^{m_i} \log p(\alpha_{ij}, \mathbb{A}_{ij}, \epsilon_{ij}, \mathbb{E}_{ij}, \mathbb{D}_i^* |\, \boldsymbol{b}_i, \boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{b}_i^T \boldsymbol{\Sigma}^{-1}\boldsymbol{b}_i\Big) = \sum_{i=1}^{n} \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i)$$

Denote the Q-function $\ell^{\mathsf{Q}}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \mathbb{E}\, \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i)$ where the expectation is over the conditional distribution of $\boldsymbol{b}_i$ given the observations at the current parameter estimates. We solve the following maximization problem involving a penalized Q-function for variable selection:

$$\begin{aligned} &\max_{\boldsymbol{\theta},\, \boldsymbol{\Sigma}\succ 0} \;\; \ell^{\mathsf{Q}}(\boldsymbol{\Theta}) - n\lambda \sum_{s=1}^{5}\sum_{r=1}^{p}\Big(c_{sr}|\,\beta_{sr}| + d_{sr}\, \boldsymbol{\Sigma}_{rr}^{(s)}\, \mathbb{I}\{r \in \mathcal{I}_c\}\Big) \\ &\text{subject to } \mathfrak{f}^{(s)}(\boldsymbol{\beta}^{(s)}) \le 0,\; s = 1, \cdots, 5\ . \end{aligned} \tag{9}$$

Here, $\boldsymbol{\beta}^{(s)} = \{\beta_{sr} : r \in \mathcal{I}\}$ and $\boldsymbol{\Sigma}$ is notationally generalized to include random effects corresponding to all p fixed effects – time invariant or not by introducing harmless zero rows and columns corresponding to time-invariant effects. This is done for presentational ease only to keep the indices same for the fixed and random effects and such degenerate large $\mathbf{\Sigma}$ matrix never crops in the computations. Also, $\mathbf{\Sigma}^{(s)}_{rr}$ is the r^{th} element of the vector $\left(\mathbf{\Sigma}_{1+p_c(s-1),1+p_c(s-1)}, \ldots, \mathbf{\Sigma}_{p_c s, p_c s}\right)$ which represents the segmented covariance matrix corresponding to the sth model, $\mathcal{I}_c$ is the index set of all composite effects and $\lambda > 0$ is the common regularization parameter which is chosen using a BIC-type criterion (Bondell et al., 2010, Lin et al., 2013, Hui et al., 2017a) given by $\texttt{BIC}_\lambda = -2\ell^{\mathsf{Q}}(\widehat{\mathbf{\Theta}}) + \log(n)\text{dim}(\widehat{\mathbf{\Theta}})$ where $\text{dim}(\widehat{\mathbf{\Theta}})$ is the number of non-zero components in $\widehat{\mathbf{\Theta}}$.

In many practical applications the composite effects impose the following hierarchy between fixed and random effects: a random component can have a non-zero coefficient only if its corresponding fixed effect is non-zero (Hui et al., 2017a). To induce such hierarchical selection of fixed and random effects, we solve equation (9) using a re-weighted ℓ_1 minimization algorithm that alternates between estimating $\mathbf{\Theta}$ and redefining the data-driven weights $(c_{sr}, d_{sr}) \in \mathbf{R}^2_+$ such that the weights used in any iteration are computed from the solutions of the previous iteration and are designed to maintain the hierarchy in selecting the fixed and random effects through their construction (see Candes et al. (2008), Zhao and Kočvara (2015), Lu et al. (2015) for details on these kind of approaches). Suppose $\Theta^{(t)}$ denote the solution to the maximization problem in equation (9) at iteration t. Then we set $c^{(t)}_{sr} = \min\left(|\beta^{(t)}_{sr}|^{-\nu}, \epsilon_1^{-1}\right)$ and $d^{(t)}_{sr} = \min\left(|\mathbf{\Sigma}^{(s,t)}_{rr}|^{-\nu}|\beta^{(t)}_{sr}|^{-\nu}, \epsilon_2^{-1}\right)$ for iteration $(t+1)$ with $\nu = 2$. We take $\epsilon_1 = 10^{-2}$ to provide numerical stability and to allow a non-zero estimate in the next iteration given a zero valued estimate in the current iteration (Candes et al., 2008) and fix $\epsilon_2 = 10^{-4}$ to enforce a large penalty on the corresponding diagonal element of $\mathbf{\Sigma}$ in iteration $(t+1)$ whenever $|\beta^{(t)}_{sr}| = 0$. Note that whenever $r \in \mathcal{I}_c$, the penalty d_{sr} on the diagonal elements of $\mathbf{\Sigma}$ encourages hierarchical selection of random effects. In section C.2 of the supplementary material we conduct simulation experiments to demonstrate this property of our re-weighted ℓ_1 procedure for solving equation (9). We end this section with the observation that although the maximization problem based on criterion (9) does not conduct any selection on the association parameters $\boldsymbol{\eta}$, it achieves that goal implicitly through the selection of the random effects.

5 Estimation procedure

In this section, we discuss two key aspects of the estimation process.

Solving the maximization problem - We use an iterative algorithm to solve the maximization problem in equation (9) which is analogous to the monte carlo EM (MCEM) algorithm of Wei and Tanner (1990). Let $\boldsymbol{\Theta}^{(t)}$ denote the parameter estimates at iteration t. In iteration $t+1$, the MCEM algorithm performs the following two steps until convergence:

E-step Recall $\mathbb{Y}_i = [\mathbb{Y}_i^{(s)}, s = 1, \ldots, 5]$. Evaluate $\ell_{(t)}^{\mathsf{Q}}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{b}_i|\boldsymbol{\Theta}^{(t)}, \mathbb{Y}_i}\ \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i)$ where the expectation above is taken with respect to the conditional distribution of $\boldsymbol{b}_i$ given the observations $\mathbb{Y}_i$ at the current estimates $\boldsymbol{\Theta}^{(t)}$. Thus,

$$\begin{aligned} \mathrm{E}_{\boldsymbol{b}_i|\boldsymbol{\Theta}^{(t)}, \mathbb{Y}_i} \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \mathbf{b}_i) &= \int \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i)\ p(\boldsymbol{b}_i \mid \alpha_i, \mathbb{A}_i, \epsilon_i, \mathbb{E}_i, \mathbb{D}_i^*, \boldsymbol{\Theta}^{(t)})\, d\boldsymbol{b}_i \\ = \exp\{-\ell_i(\boldsymbol{\Theta}^{(t)})\} & \int \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i)\ p(\alpha_i, \mathbb{A}_i, \epsilon_i, \mathbb{E}_i, \mathbb{D}_i^* |\ \boldsymbol{\theta}^{(t)}, \boldsymbol{b}_i,)\ \phi_{p_c}(\boldsymbol{b}_i | \mathbf{0}, \boldsymbol{\Sigma}^{(t)})\, d\boldsymbol{b}_i \end{aligned}$$

where, $\phi_{p_c}(\,\cdot\,|\,\mathbf{0}, \boldsymbol{\Sigma}^{(t)})$ is the p_c dimensional normal density with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}^{(t)}$. Note that the expectation involves a multivariate integration with respect to the random effects $\boldsymbol{b}_i$ which is evaluated by Monte Carlo integration. We approximate it as:

$$\Big(\sum_{d=1}^{D} \ell_i^{\mathsf{cl}}(\boldsymbol{\Theta}, \boldsymbol{b}_i^d)\ p(\alpha_i, \mathbb{A}_i, \epsilon_i, \mathbb{E}_i, \mathbb{D}_i^* |\ \boldsymbol{\theta}^{(t)}, \boldsymbol{b}_i^d)\Big) \Big/ \Big(\sum_{d=1}^{D} p(\alpha_i, \mathbb{A}_i, \epsilon_i, \mathbb{E}_i, \mathbb{D}_i^* |\ \boldsymbol{\theta}^{(t)}, \boldsymbol{b}_i^d)\Big)$$

where $\boldsymbol{b}_i^d$ is a random sample from $\phi_{p_c}(\,\cdot\,|\,\mathbf{0}, \boldsymbol{\Sigma}^{(t)})$ and $D = 2000$ is the number of monte carlo samples.

M-step Solve the following maximization problem with data driven adaptive weights $(c_{sr}^{(t)}, d_{sr}^{(t)})$

$$\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\theta}, \boldsymbol{\Sigma} \succ 0}{\arg\max}\ \ \ell_{(t)}^{\mathsf{Q}}(\boldsymbol{\Theta}) - n\lambda \sum_{s=1}^{5} \sum_{r=1}^{p} \Big(\, c_{sr}^{(t)} |\, \beta_{sr}| +\ d_{sr}^{(t)}\, \boldsymbol{\Sigma}_{rr}^{(s)}\, \mathbb{I}\{r \in \mathcal{I}_c\}\Big)$$
$$\text{subject to } \mathfrak{f}^{(s)}(\boldsymbol{\beta}^{(s)}) \leq 0,\ s = 1, \cdots, 5\ .$$

The maximization problem above decouples into separate components that estimate $\boldsymbol{\beta}^{(s)}$ as solutions to convex problems and $\boldsymbol{\Sigma}$ as a solution to a non-convex problem. To solve the convex problems involving $\boldsymbol{\beta}^{(s)}$, we use a proximal gradient descent algorithm after reducing the original problem to an ℓ_1 penalized least squares fit with convex constraints. See James et al. (2013) for related approaches of this kind. For estimating $\mathbf{\Sigma}$, we use the coordinate descent algorithm of Wang (2014) that solves a lasso problem and updates $\mathbf{\Sigma}$ one column and row at a time while keeping the rest fixed. Further details regarding our estimation procedure is presented in section A of the supplementary material.

Split and Conquer - To enhance the computational efficiency of the estimation procedure, we use the split-and-conquer approach of Chen and Xie (2014) to split the full set of n players into K non-overlapping groups and conduct variable selection separately in each group by solving K parallel maximization problems represented by equation (9). Following Chen and Xie (2014), the selected fixed and random effects are then determined using a majority voting scheme across all the K groups as described below.

Let $\widehat{\boldsymbol{\beta}}^{(s)}[k]$ and $\widehat{\mathbf{\Sigma}}^{(s)}[k]$ denote, respectively, the estimate of the fixed effect coefficients for model s and the estimate of the p_c diagonal elements of $\mathbf{\Sigma}$ for model s on split k obtained by solving the maximization problem (9), where $k = 1, \ldots, K$. We construct the set of selected effects as:

$$\textsf{Set of Fixed Effects:}\quad \widehat{\mathcal{I}}^{(s)} = \Big\{r\colon \sum_{k=1}^{K} \mathbb{I}(\widehat{\beta}_{sr}[k] \neq 0) \geq w_0,\ r = 1, \ldots, p\Big\}$$

$$\textsf{Set of Random Effects:}\quad \widehat{\mathcal{I}}_R^{(s)} = \Big\{r\colon \sum_{k=1}^{K} \mathbb{I}(\widehat{\mathbf{\Sigma}}^{(k)}_{r+p_c(s-1), r+p_c(s-1)} > 0) \geq w_1,\ r = 1, \ldots, p_c\Big\}$$

Here $w_0,\ w_1$ are pre-specified thresholds determining the severity of the majority voting scheme. For large datasets as in mobile apps application, a distributed computing framework utilizing the above scheme leads to substantial reduction in computation time. Section C of the supplementary material presents a discussion of the split-and-conquer approach along with numerical experiments that demonstrate the applicability of this method in our setting where data driven adaptive weights are used in the penalty and variable selection is conducted simultaneously across multiple models. Finally, based on the selected fixed and random effect components in $\widehat{\mathcal{I}}^{(s)}$ and $\widehat{\mathcal{I}}_R^{(s)}$, we use the entire data and estimate their effects more accurately by maximizing the likelihood based on only those components using the standard EM algorithm.

6 Analysis of freemium mobile games using CEZIJ

We apply our proposed CEZIJ methodology to the freemium mobile game data discussed in Section 2. This dataset holds player level gaming information for 38,860 players observed over a period of 60 days. The analyses presented here uses a sample of 33,860 players for estimation and the remaining $5,000$ players for out of sample validation. See section D in the supplementary material for a detailed description of the data. For sub-models $s = 1, \ldots, 4$, we consider a set of 30 predictors, of which 24 can have composite effects. The 24 composite effects are listed in table 3 (Serial No 1-24) of the supplementary material. The remaining 6 predictors are the 6 promotion strategies summarized in section D and table 5 of the supplementary material. We treat these promotion strategies as potential fixed effects with no corresponding random effect counterparts. For the dropout model, which shares its random effects with the four sub-models, the entire list of 30 candidate predictors is taken as potential fixed effects. Overall, the selection mechanism must select random effects from a set of 100 potential random effects (24 for each of the four sub-models and their 4 intercepts) and select fixed effects from a set of 155 potential fixed effects (30 for each of the five sub-models and their 5 intercepts).

To model the responses at time point j, we consider time $j-1$ values of the predictors that contain gaming characteristics of a player simply because at time j these characteristics are known only upto the previous time point $j-1$. These gaming characteristics are marked with an $(*)$ in table 3 of the supplementary material. All the remaining predictors like weekend indicator and the 6 indicator variables corresponding to the promotion strategies are applied at time j. We initialize the CEZIJ algorithm by fitting a saturated model on a subset of 200 players, which was also used to initialize the weights $c_{sr},\ d_{sr}$ in criterion (9). Finally, our algorithm is run on $K = 20$ splits where each split holds $n_k = 1693$ randomly selected players with the majority voting parameters $w_0,\ w_1$ fixed at 12 and the regularization parameter λ_k is chosen as that value of $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.25, 0.5, 1, 5\}$ which minimizes $\texttt{BIC}_\lambda$. Table 6 in section E of the supplementary material presents the voting results for each candidate predictor across the 20 splits.

6.1 The fitted joint model and its interpretations

The final list of selected predictors and their estimated fixed effect coefficients for the sub-models of Activity Indicator (AI), Activity time (daily total time played), Engagement Indicator (EI), Engagement amount and Dropout is presented in Table 2. See Table 3 (section D of the supplementary material) for the description of the covariates. The selected composite effects are those predictors that exhibit a $(*)$ over their coefficient estimates in Table 2. All the selected fixed and random effects obey the hierarchical structure discussed in section 4. We next discuss the fitted coefficients for each sub-model.

Activity Indicator - For modeling probability of AI, the CEZIJ methodology selects 18 fixed effects of which 14 are composite effects. As AI forms the base of our joint model, the fixed effects of its estimated marginal distribution have the least nuanced interpretation among the 5 sub-models. All other things remaining constant, there is an overall increase in the odds for AI by 35% on the weekends and an 8% increase in odds for each level advancement in the game. Similarly, the conditional odds is boosted by 20%, 14% and 9% respectively if the gacha game was played or robot purchases or upgrades were made in the previous days. Absence of log-in in the previous day adversely affects the odds with an average decrement of 88% for each absent day. Promotions II and VI, which provide sale on robots at different dosages are positively associated and increase odds of AI by approximately 20% and 30% respectively.

Activity Time - In this case the selection mechanism selects 17 fixed effects of which 15 are composite effects. The signs on the coefficients of `timesince` and `weekend` align with the constraints imposed on them and along with the game characteristics like number of primary and auxiliary fights played, level progressions, and robot upgrades, continue to provide a similar interpretation as with the AI model. This is the second layer of joint model which is conditioned on positive login occurrence. A key difference between these two models, however, lies in the inclusion of predictors `avg_session_length`, `gacha_sink` and `pfight_source`. They indicate that, keeping other things fixed, players interacting with the game through spending in-game currencies or winning the same through principal fights on the previous day have the natural incentive to spend more in-game time on the following day. In line with the monotonicity constraints imposed on the promotion strategies for this model, the coefficient for promotion VI is both positive and bigger than the coefficient for promotion II thus indicating that the strategy to promote sale on all robots has a higher impact on activity time than the strategy to offer the special particular 'Boss' robots at a discount.

Engagement Indicator and Amount - Recall from section 3 that we use an EZI Log Normal model for the engagement amount by first building a separate model for the probability of EI given activity. For the sub-models that model EI and the engagement amount, the CEZIJ methodology selects respectively, 22 fixed effects of which 16 are composite effects and 6 fixed effects of which 2 are composite effects. Direct interpretation of the fixed effect coefficients is difficult here, as this sub-model is conditioned on the first two sub-models. We see that some of the key player engagement characteristics like number of auxiliary fights played, level progression, in-app virtual currency spent and earned seem to positively impact the conditional likelihood of positive engagement at subsequent time points. A significant finding is that among the three different fight modes, only auxiliary fight second mode which involve time restricted fights seems to lead to substantially higher player engagement implying that all other variables remaining constant, player engagement in game promotion through social media is more while playing time attack fights.

Dropout - In this case, the selection mechanism selects 9 fixed effects. The sign on the coefficient for `timesince` is positive, which is natural, and indicates that players who do not frequent the game often (low frequency of AI) exhibit a high likelihood of dropping out at subsequent time points. It is also interesting to see, through `gacha_sink`, that all else being equal, players who spend more of their virtual currencies on gacha exhibit a high likelihood of dropping out at subsequent time points. This can potentially be explained through a `"make-gacha-work-for-all-players"`(Agelle, 2016) phenomenon where the player spends a major portion of her virtual currency on gacha however the value of the items won is largely worthless when compared to the amount of currency spent, thus inducing a lack of interest in the game at future time points. All the promotions with exception of promotion V, reduces the odds of dropouts validating their usage as retention

Table 2: Selected fixed effect coefficients and their estimates under the sub-models Act. Indicator, Activity Time, Engag. Indicator and Engagement Amount and Dropout. The selected random effects are those variables that exhibit a ($*$) over their estimates. See Table 3 for a detailed description of the covariates.

Predictors	Act. Indicator	Activity Time	Engag. Indicator	Engag. Amount	Churn
	$\widehat{\boldsymbol{\beta}}^{(1)}$	$\widehat{\boldsymbol{\beta}}^{(2)}$	$\widehat{\boldsymbol{\beta}}^{(3)}$	$\widehat{\boldsymbol{\beta}}^{(4)}$	$\widehat{\boldsymbol{\beta}}^{(5)}$
intercept	-4.648*	0.932*	-1.560*	0.953*	-1.902
avg_session_length	–	0.269*	0.198*	–	–
p_fights	0.378*	0.169*	-0.126*	–	–
a1_fights	0.303*	0.379*	0.274*	–	–
a2_fights	0.334*	0.216*	-0.492*	–	–
level	0.084*	0.304*	0.282*	–	–
robot_played	–	–	–	–	–
gacha_sink	–	0.201*	0.509*	–	0.129
gacha_premium_sink	–	–	–	–	–
pfight_source	–	0.144*	–	–	–
a1fight_source	0.030	-0.239*	-0.727*	–	–
a2fight_source	-0.240*	-0.192*	0.482*	0.331*	–
gacha_source	0.182*	-0.212*	–	–	–
gacha_premium_source	–	–	0.133*	–	–
robot_purchase_count	0.134*	–	–	–	–
upgrade_count	0.093*	0.112*	0.404*	–	–
lucky_draw_wg	–	–	-0.240*	–	–
timesince	-2.065*	-0.641*	-0.229*	–	3.502
lucky_draw_og	-0.230*	–	-0.469*	–	–
fancy_sink	–	–	-0.110*	–	–
upgrade_sink	0.037*	–	-0.272*	–	–
robot_buy_sink	–	–	0.159	–	–
gain_gachaprem	–	–	–	–	–
gain_gachagrind	-0.127*	0.180*	–	–	–
weekend	0.302*	0.358*	–	–	–
promotion I	–	–	–	-1.153	-0.894
promotion II	0.178	0.134	-0.189	–	-0.934
promotion III	-0.129	–	-0.166	-1.791	-3.500
promotion IV	–	–	0.164	-3.345	-0.673
promotion V	–	–	-5.000	–	0.828
promotion VI	0.290	0.249	0.131	-2.389	-1.509

schemes.

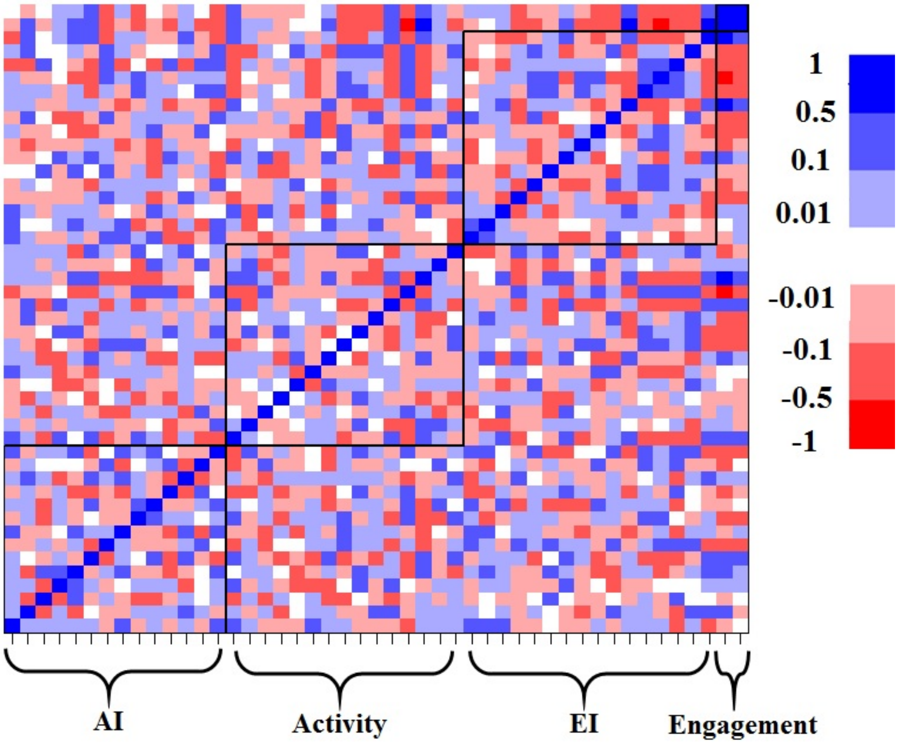


Figure 4: Heatmap of the 47 × 47 correlation matrix obtained from $\widehat{\boldsymbol{\Sigma}}$. On the horizontal axis are the selected composite effects of the four sub-models: AI, Activity Time, EI and Engagement Amount. The horizontal axis begins with the `intercept` from the AI model and ends with `a2fight_source` from the Engagement Amount model.

From the heatmap in figure 4, the random effects of the selected composite effect predictors demonstrate correlations within the four sub-models that were modeled jointly, indicating that players exhibit idiosyncratic profiles over time. Moreover, we notice several instances of cross correlations across the four sub-models. For example from figure 5, the random effect associated with the number of championship fights played (predictor `p_fights`) in the AI model has a positive correlation with the amount of virtual currency earned through auxiliary fights (predictor `a2_fights_source`) played in the model for Activity Time which suggests that the modeled responses are correlated for a player. Our

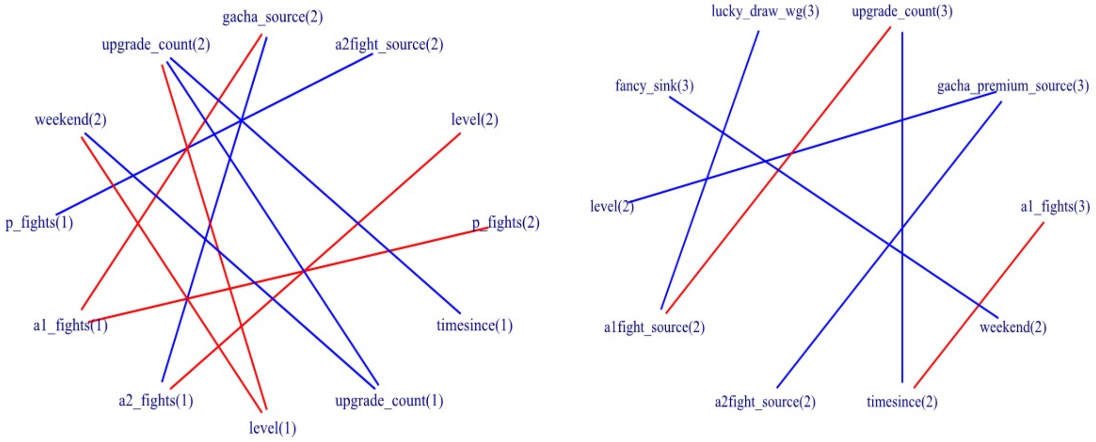


Figure 5: Two networks that demonstrate several cross correlations across the models. Blue line represents positive correlation and red line represents negative correlation. The model numbers are inside the parenthesis next to the predictor names. Left: Key cross correlations between the sub-models AI and Activity Time. Right: Key cross correlations between the sub-models Activity Time and EI.

joint model allows us to borrow information across these related responses and may aid game managers and marketers in understanding how the outcomes depend on each other.

6.2 Out of sample validation

We use the hold-out sample of $5,000$ players from the original data for assessing the predictive accuracy of our model. Our scheme consists of predicting the four outcomes - AI, activity time, EI and engagement amount, dynamically over the next 29 days using the fitted model discussed in section 6.1. Note that the time frame of prediction covers the first 30 days of game usage for each player, and so by definition, no player drops out which leaves us with the aforementioned four outcomes to predict. As benchmarks to our fitted model, we consider four competing models - Benchmark I to Benchmark IV which we describe below.

For Benchmark I we consider a setup where there are no random effects, the outcomes are not modeled jointly and variable selection is conducted using the R-package `glmmLasso` (Schelldorfer et al., 2014) that uses an ℓ_1-penalized algorithm for fitting high-dimensional

Table 3: Results of predictive performance of CEZIJ model and Benchmarks I to IV. For activity and engagement indicators, the false positive (FP) rate / the false negative (FN) rate averaged over the 29 time points are reported. For non-zero activity time and engagement amounts, the ratio of prediction errors (10) of Benchmarks I to IV to CEZIJ model averaged over the 29 time points are reported.

Sub-model	Benchmark I	Benchmark II	Benchmark III	Benchmark IV	CEZIJ
Activity Indicator	1.32% / 6.71%	0.27% / 7.83%	1.19% / 6.33%	5.92% / 4.15%	5.86% / 4.12%
Total Time Played	1.742	1.961	4.662	1.041	1
Engagement Indicator	0.09% / 1.87%	0% / 1.89%	0.05% / 1.89%	3.56% / 1.48%	3.54% / 1.47%
Engagement Amount	1.408	8.619	1.217	1.067	1

generalized linear mixed models (GLMMs) with logit links for AI, EI and identity link for the two continuous outcomes of positive activity time and engagement amount. In case of Benchmark II, we continue to model the outcomes separately and use the R-package `rpql` (Hui et al., 2017b) that performs joint selection of fixed and random effects in GLMMs using a regularized PQL (Breslow and Clayton, 1993) with similar link functions as used in Benchmark I. The remaining two Benchmark models rely on the selected variables from the CEZIJ model itself and do not conduct their respective variable selection. In particular, Benchmark III uses the selected predictors from the CEZIJ methodology and models the outcomes via generalized linear models with logit links for AI, EI and identity link for the two continuous outcomes of positive activity time and engagement amount. Thus Benchmark III, like Benchmark I, represents a setup where there are no random effects and the outcomes are not modeled jointly. Benchmark IV, on the other hand, represents a more sophisticated setup wherein it resembles the fitted CEZIJ model in every aspect except that the random effects across the four sub-models are not correlated. It achieves this by using the selected fixed and composite effects from CEZIJ model but employs a slightly modified covariance matrix $\breve{\mathbf{\Sigma}}$ where the covariances between random effects originating from the different sub-models are set to 0, thus representing a setup where the outcomes are not modeled jointly.

The out of sample validation requires predicting the responses dynamically over time. For Benchmarks I and III this step is easily carried out by running the fitted model on the validation data. However, for Benchmark II, IV and CEZIJ model the prediction mechanism must, respectively, estimate the latent random effects and appropriately account for the endogenous nature of the responses. To do that we utilize the simulation scheme discussed in section 7.2 of Rizopoulos (2012) and section 3 of Rizopoulos (2011), and calculate the expected time j responses given the observed responses until time $j-1$, the estimated parameters and the event that the player has not churned until time $j-1$ (details provided in section B of the supplementary material). Table 3 summarizes the results of predictive performance of CEZIJ and the benchmark models. For AI and EI, table 3 presents, for each model, the false positive (FP) rate and the false negative (FN) rate respectively averaged over the 29 time points. The FP rate measures the percentage of cases where the model incorrectly predicted activity (or engagement) whereas the FN rate measures the percentage of cases where the model incorrectly predicted no activity (or no engagement). Benchmark II, for example, exhibits the lowest FP rate and has the highest FN rate followed by Benchmark III. The low FP rate of Benchmark II, however, belies the relatively poor performance of this model in predicting zero inflated responses which becomes apparent in the higher FN rates especially for the EI model. The CEZIJ model alongwith Benchmark IV, on the other hand, have the lowest FN rates demonstrating their relatively superior ability in predicting the zero inflated responses of AI and EI. For positive activity times and positive engagement values we take a slightly different approach and first calculate the time j prediction errors PE_j for the Benchmark models and CEZIJ as follows. For any model $\mathcal{M} \in \{$Benchmark I, ..., Benchmark IV, CEZIJ$\}$, we define $\mathrm{PE}_j^{\mathcal{M}}$ for sub-model $s=2$ at time $j=1,\ldots,29$ as

$$\mathrm{PE}_j^{\mathcal{M}}(\mathbb{Y}^{*(s)}, \widehat{\mathbb{Y}}^{*(s)}) = \sum_{i=1}^{n} \Big| \log \mathbb{Y}_{ij}^{*(s)} - \log \widehat{\mathbb{Y}}_{ij}^{*(s)} \Big| \tag{10}$$

where $\mathbb{Y}_{ij}^{*(s)} = \mathbb{Y}_{ij}^{(s)}$ if $\alpha_{ij}=1$ and 1 otherwise, and $\widehat{\mathbb{Y}}_{ij}^{*(s)} = \widehat{\mathbb{Y}}_{ij}^{(s)}$ if $\widehat{\alpha}_{ij}=1$ and 1 otherwise with $\widehat{\mathbb{Y}}_{ij}^{(s)}$, $\widehat{\alpha}_{ij}$ being model $\mathcal{M}$ predictions of activity time, AI, respectively, for player i at time j. The time j prediction error for sub-model $s=4$ is also defined in a similar fashion with α_{ij}, $\widehat{\alpha}_{ij}$ replaced with ϵ_{ij}, $\widehat{\epsilon}_{ij}$ respectively and measures the total absolute deviation of the prediction from the truth at any time j. For notational convenience the dependence of $\mathrm{PE}_j^{\mathcal{M}}$ on α_{ij}, $\widehat{\alpha}_{ij}$ (or ϵ_{ij}, $\widehat{\epsilon}_{ij}$) have been suppressed but the inclusion of these predicted and observed indicators in equation (10) is aimed at exploiting the dependencies between the responses, if any. For the two sub-models ($s = 2, 4$) table 3 presents the ratio of the prediction errors of the Benchmarks to the CEZIJ model averaged over the 29 time points where a ratio in excess of 1 indicates a larger absolute deviation of the prediction from the truth when compared to CEZIJ model. All Benchmark models exhibit prediction error ratios bigger than 1 with Benchmarks II and III being the worse for engagement amount and activity time models respectively. Benchmark IV, on the other hand, profits from the structure of the various components of CEZIJ model but is unable to account for the dependencies between the responses which is reflected in its prediction error ratios being slightly bigger than 1 but alongwith the CEZIJ model, it continues to demonstrate superior prediction error ratios across the two sub-models.

6.3 Player segmentation using predicted churn probabilities

Player sub-populations with similar churn characteristics over time provide valuable insights into user profiles that are more likely to dropout and can be used to design future retention policies specifically targeting those characteristics. In this section we use the fitted churn model of section 6.1 to predict the temporal trajectories of churn probabilities on a sample of $1,000$ players who are 30 days into the game and use the predicted probabilities over the next 25 days to cluster the players into homogeneous sub-groups. The churn probabilities are predicted in a similar fashion as discussed in section 6.2 and section B of the supplementary material where the churn probability at time j is predicted conditional on the estimated parameters, the observed responses until time $j - 1$ and the event that the player has not churned until time $j - 1$. To determine the player subgroups, we use `R` package `fda.usc` to cluster the rows of the 1000×25 predicted churn probability matrix using functional K-means clustering. We use the prediction strength algorithm of Tibshirani and Walther (2005) to determine the number of clusters.

In figure 6 the three cluster centroids segment the sample into groups which demonstrate distinct temporal churn profiles. For instance, cluster 3, which holds almost 48% of the players, exhibits rising churn probabilities until day 5 but tapers down under the influence

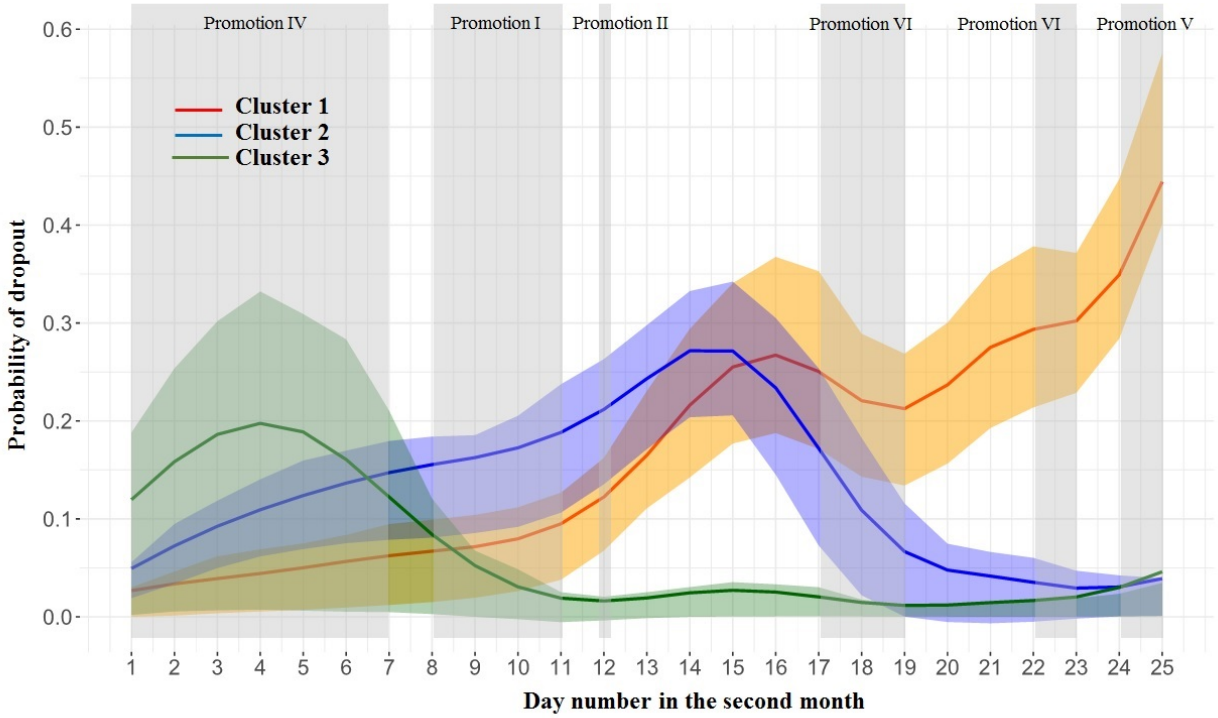


Figure 6: Functional cluster analysis of predicted dropout probabilities over time. The plot presents three cluster centroids. The shaded band around the centroids are the 25^{th} and 75^{th} percentiles of the churn probabilities. The vertical shaded regions in the graph correspond to the days on which different promotion strategies were on effect. The number of clusters were identified using prediction strength (Tibshirani and Walther, 2005).

promotions I, II and VI. Cluster 2, with 34% of the players, has a different trajectory than cluster 3 and appears to respond favorably to promotion VI. Of particular importance are those players that belong to cluster 1 which holds 18% of the players and is characterized by rising churn probabilities over time. The churn profile of this cluster represents players who have been relatively inactive in the game and continue to do so even under the influence of various promotion strategies. During days 17 to 19, their churn probabilities are predicted to diminish under the effect of promotion VI however subsequent promotions do not appear to have any favorable impact. These segment curves suggest that there are some key differences in customer attrition patterns. For example, Cluster 1 shows increasing attrition rates over time, which suggests that the game is not able to retain these players. Cluster 2 shows increasing attrition initially, but then the attrition rate starts to decline significantly after 45 days. This segment is potentially beneficial to the platform as it demonstrates that there is a core set of players who are loyal to the game. Players in Cluster 3 on average start with a much higher attrition rate than the other two segments, but their attrition rate tapers down significantly after five days and then stays at a very low level over time. Interestingly, Cluster 3 seems to be responding to promotions I, II and IV. These differences in user behavior across the segments can be leveraged to increase the efficiency of player retention policies. They also suggest that the platform should adopt different business strategies. For instance, in Cluster 3 many players have been weeded out early. This indicates that short term visitors to the gaming portal have left the platform more quickly in Cluster 3 compared to the other two segments. So it is important for the platform to emphasize promotional activities that increase player engagement. On the contrary, for Clusters 1 and 2, it is important for the platform to emphasize promotional activities that increase player log-in or activity. This relative emphasis across the segments can increase the efficiency of marketing campaigns.

7 Discussion

We propose a very scalable joint modeling framework CEZIJ for unified inference and prediction of player activity and engagement in freemium mobile games. The rapid growth of mobile games globally has generated significant research interest in different business areas such as marketing, management and information sciences. Our proposed algorithm conducts variable selection by maintaining the hierarchical congruity of the fixed and random effects and produces models with interpretable composite effects. A key feature of our framework is that it allows incorporation of side information and domain expertise through convexity constraints. We exhibit the superior performance of CEZIJ in producing dynamic predictions. It is also used to segment players based on their churn rates, with the analysis revealing several idiosyncratic player behaviors that can be used for targeted marketing of players in future freemium games. The segmentation findings have important business implications for monetization of the platforms. They can be used to enhance the effectiveness and efficiency of promotional activities and also future user acquisition and retention strategies.

Our inferential framework is based on modern optimization techniques and is very flexible. It can be used in a wide range of big-data applications that need analyzing multiple high-dimensional longitudinal outcomes along with a time-to-event analysis. In future, we would like to extend our joint modeling program for providing comprehensive statistical guidance regarding the growth, development and optimal pricing of generic digital products that use the freemium model. For that purpose, it will be interesting to investigate extensions of our CEZIJ modeling framework, in particular, the possibility of incorporating non-parametric components for modeling the nonlinear time effects since player behavior may change over time. Furthermore, the current dropout model in equation (8) may be enhanced to include more sophisticated structures involving cumulative effects parametrization and conduct variable selection on the high dimensional vector of association parameter $\boldsymbol{\eta}$, which the current CEZIJ framework implicitly achieves through the selection of the random effects. An alternative and computationally less demanding approach may be to consider the following low dimensional representation wherein the dropout model is of the form $\mathsf{logit}(\lambda_{ij}) = \boldsymbol{x}_{ij}^{(5)T}\boldsymbol{\beta}^{(5)} + \sum_{s=1}^{4}\eta_s \boldsymbol{z}_{ij}^{(s)T}\boldsymbol{b}_i^{(s)}$ so that $\boldsymbol{\eta}$ is then only a 4×1 vector. Finally, while the focus of this paper is the CEZIJ modeling framework and its applicability in the disciplined study of freemium behavior and other applications that needs analyzing multiple high-dimensional longitudinal outcomes along with a time-to-event analysis, a natural extension of our work, as future research, will be targeted towards estimating standard errors of the estimated coefficients and confidence intervals under the CEZIJ framework using ideas from recent developments in post-selection inference (see Javanmard and Montanari (2014), Lee et al. (2016) for example).

Of the thousands of freemium games that are developed every month, very few of them go on to make adequate amount through IAP (in-app purchases). Most games resemble our data where a significant part of the revenue is earned through in-game ads and social media usages. In these games, such low incidence of real money purchases present a challenge in model development as the robustness of the estimated model coefficients will be significantly impacted in case real money purchases are modeled as a separate response variable. Thus, in very low IAP incidence games it is useful to model the combined revenue using game specific weights to blend direct and indirect engagement as is done in this paper. For games with significant amount of IAP, we envision modeling direct and indirect engagement separately and study their interactions.

8 Acknowledgments

We thank Professors Gil Appel and Milan Miric for helpful discussions and the Editor, the Associate Editor and two anonymous referees for several constructive suggestions that have greatly improved the presentation of the paper.

References

Agelle, P. (2016). Getting gacha right: Tips for creating successful in-game lotteries. *PocketGamer*. Available at http://www.pocketgamer.biz/comment-and-opinion/63620/getting-gacha-right-tips-for-creating-successful-in-game-lotteries/.

Alfò, M., Maruotti, A., and Trovato, G. (2011). A finite mixture model for multivariate counts under endogenous selectivity. *Statistics and Computing*, 21(2):185–202.

AppBrain (2017). Free vs. paid android apps. *AppBrain, July 24*. Available at http://www.appbrain.com/stats/free-and-paid-android-applications.

Appel, G., Libai, B., Muller, E., and Shachar, R. (2017). Retention and the monetization of apps.

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

Banerjee, T., Mukherjee, G., and Sun, W. (2018). Adaptive sparse estimation with side information. *arXiv preprint arXiv:1811.11930*.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077.

Boudreau, K., Jeppesen, L. B., and Miric, M. (2017). Freemium, network effects and digital competition: Evidence from the introduction of game center on the apple appstore.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.

Candes, E. J., Wakin, M. B., and Boyd, S. P. (2008). Enhancing sparsity by reweighted 1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905.

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

Diele, O. (2013). State of online gaming report. *Spil Games*. Available at http://auth-83051f68-ec6c-44e0-afe5-bd8902acff57.cdn.spilcloud.com/v1/archives/1384952861.25_State_of_Gaming_2013_US_FINAL.pdf.

Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. *Annals of statistics*, 40(4):2043.

Garg, R. and Telang, R. (2012). Inferring app demand from publicly available data.

Greene, W. (2009). Models for count data with endogenous participation. *Empirical Economics*, 36(1):133–173.

Guo, X. and Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58(1):16–24.

Gustafsson, A., Johnson, M. D., and Roos, I. (2005). The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention. *Journal of marketing*, 69(4):210–218.

Haans, H., Raassens, N., and van Hout, R. (2013). Search engine advertisements: The impact of advertising statements on click-through and conversion rates. *Marketing Letters*, 24(2):151–163.

Han, C. and Kronmal, R. (2006). Two-part models for analysis of agatston scores with possible proportionality constraints. *Communications in StatisticsTheory and Methods*, 35(1):99–111.

Hatfield, L. A., Boye, M. E., Hackshaw, M. D., and Carlin, B. P. (2012). Multilevel bayesian models for survival times and longitudinal patient-reported outcomes with many zeros. *Journal of the American Statistical Association*, 107(499):875–885.

Hui, F. K., Müller, S., and Welsh, A. (2017a). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statistica Sinica*, 27(2).

Hui, F. K., Müller, S., and Welsh, A. (2017b). Joint selection in mixed models using regularized pql. *Journal of the American Statistical Association*, 112(519):1323–1333.

Hui, S. K., Inman, J. J., Huang, Y., and Suher, J. (2013). The effect of in-store travel distance on unplanned spending: Applications to mobile promotion strategies. *Journal of Marketing*, 77(2):1–16.

Hwong, C. (2016). Using audience measurement data to boost user acquisition and engagement. *Verto Analytics*. Available at http://www.vertoanalytics.com/report-leveling-mobile-game-using-audience-measurement-data-boost-user-acquisition-engagement/.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503.

James, G. M., Paulson, C., and Rusmevichientong, P. (2013). Penalized and constrained regression. *Technical Report*.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

Jerath, K., Fader, P. S., and Hardie, B. G. (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. et al. (2013). On statistics, computation and scalability. *Bernoulli*, 19(4):1378–1390.

Jordan, M. I., Lee, J. D., and Yang, Y. (2018). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, (just-accepted).

Kanerva, T. (2016). Cultures combined: Japanese gachas are sweeping f2p mobile games in the west. *GameRefinery*. Available at http://www.gamerefinery.com/japanese-gachas-sweeping-f2p-games-west/.

Koetsier, J. (2015). Why 2016 is the global tipping point for the mobile economy. *Tune*. Available at https://www.tune.com/blog/global-mobile-why-2016-is-the-global-tipping-point-for-the-mobile-economy/.

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

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927.

Lee, J. D., Sun, Y., Liu, Q., and Taylor, J. E. (2015). Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*.

Lin, B., Pang, Z., and Jiang, J. (2013). Fixed and random effects selection by reml and pathwise coordinate optimization. *Journal of Computational and Graphical Statistics*, 22(2):341–355.

Liu, C. Z., Au, Y. A., and Choi, H. S. (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.

Lu, C., Lin, Z., and Yan, S. (2015). Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization. *IEEE Transactions on Image Processing*, 24(2):646–654.

MarketingCharts (2017). App retention rates still low, but improving. *Marketing Charts, Feb 20.* Available at http://www.marketingcharts.com/online/app-retention-rates-still-low-but-improving-75135/.

McCulloch, C. (2008). Joint modelling of mixed outcome types using latent variables. *Statistical Methods in Medical Research*, 17(1):53–73.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170.

McDonald, E. (2017). The global games market. *Newzoo.* Available at https://newzoo.com/insights/articles/the-global-%20games-market-%20will-reach-108-%209-billion-%20in-2017-%20with-mobile-%20taking-42.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical modelling*, 5(1):1–19.

Needleman, S. E. and Loten, A. (2012). When freemium fails. *WSJ.* Available at https://www.wsj.com/articles/SB10000872396390443713704577603782317318996.

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

Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96(454):730–745.

Pan, J. and Huang, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing*, 24(5):725–738.

Peng, H. and Lu, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis*, 109:109–129.

Perro, J. (2016). Mobile apps: Whats a good retention rate? *Localytics, March 28.* Available at http://Info.Localytics.Com/Blog/Mobile-apps-Whats-A-Good-Retention-Rate.

PocketGamer (2018). Number of applications submitted per month to the itunes app store. *Pocket Gamer.* Available at http://www.pocketgamer.biz/metrics/app-store/submissions/.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., et al. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2(1):1–21.

Rizopoulos, D. (2011). Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829.

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

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

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

Schelldorfer, J., Meier, L., and Bühlmann, P. (2014). Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using 1-penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477.

Statista (2018). How many hours in a typical week would you say you play games? *Statista*. Available at https://www.statista.com/statistics/273311/time-spent-gaming-weekly-in-the-uk-by-age/.

Swrve (2016). Monetization report 2016. *swrve*. Available at https://www.swrve.com/images/uploads/whitepapers/swrve-monetization-report-2016.pdf.

Taube, A. (2013). People spend way more on purchases in free apps than they do downloading paid apps. *Business Insider, December 30*. Available at http://www.businessinsider.com/inapp-purchases-dominate-revenue-share-2013-12.

Tibshirani, R. and Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the american statistical association*, 81(393):82–86.

Toto, S. (2012). Gacha: Explaining japans top money-making social game mechanism. *Kantan Games*. Available at https://www.serkantoto.com/2012/02/21/gacha-social-games/.

Urban, G. L., Liberali, G., MacDonald, E., Bordley, R., and Hauser, J. R. (2013). Morphing banner advertising. *Marketing Science*, 33(1):27–46.

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

Wang, H. (2014). Coordinate descent algorithm for covariance graphical lasso. *Statistics and Computing*, 24(4):521–529.

Wei, G. C. and Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704.

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

Zhao, Y.-B. and Kočvara, M. (2015). A new computational method for the sparsest solutions to systems of linear equations. *SIAM Journal on Optimization*, 25(2):1110–1134.