Dataset Overview for the Paper: Predicting Daily Active Users for Match-3 Mobile Games

There are two datasets available for research and both will be considered for the empirical study. The *DNC Dataset* and the *JNC Dataset*. Both datasets have been compiled and retrieved from **Flurry Analytics**, a commercial analytics tool used by **Playlab Inc** to track user's behavior on their commercial mobile games. Some attributes were retrieved from **Google Play** such as user's ratings and daily crash reports. The *DNC* dataset refers to the Dragon Cubes game while the *JNC* dataset refers to the Jungle Cubes game. Both dataset are restricted to Android platforms only.

Here are the overall information of the datasets to be used:

Dataset Filename	Game Title	Total Downloads	Overall Rating	Timeline of Dataset
DNC Dataset Android 0511-0911	Dragon Cubes	50,000 – 100,000	4.2 out of 5	May 11, 2015 – September 11, 2015
JNC Dataset Android 0511-0911	Jungle Cubes	100,000 – 500,000	4.3 out of 5	May 11, 2015 – September 11, 2015

Both datasets spans on a similar timeline, that is May 11 – September 11,2015. A span of four months have been deemed sufficient for analysis and increasing the timespan no longer yields better results.

Definition of Attributes

Install Date	Each instance in the dataset is organized by install date. This refers to the gregorian calendar date wherein an application is installed.
Cohort Size	Refers to the total amount of users who have installed the application on the given install date.
Day X	This represents the retention of the application given a certain date and cohort size. Installation date becomes day 0. Retention rate is the percentage of returning users on a specified install date. For example, day 1 has 40.75% retention and 1200 cohort size. Therefore, 40.75% of users have managed to return on day 1 (489 users in cohort size)
CrashesANRDay1	This counts the total number of crashes and ANRs (application not responding) reports from the application. This has a negative impact for the user experience. In reality, crash reports come in **a day after** the specified install date. For example, May 11,2015 has 3 crash reports. This means that this value was only retrieved on May 12, 2015.
DailyAverageRating	This refers to the average rating by users who choose to rate the application (1 to 5, 5 being the highest) on a given date. Rating an application is not mandatory. This is a primary determination for virality. Similar to *CrashesANRDay1*, the tally comes in **a day after** the specified install date.
LevelPlayedEvents	Refers to the accumulated event tally that is triggered when a user plays a level on the application. This is triggered upon tap of the 'Play' button. This event is reported no matter the outcome of the level being played.
LevelSuccessEvents	Refers to the accumulated events that are triggered if a user **successfully completes** a level. This is triggered when the 'Win' screen is shown to the user.
LevelFailedEvents	Refers to the accumulated events that are triggered if a user **fails** a level. This is triggered when the 'Lose' screen is shown to the user.
Sessions	Refers to the total amount of play sessions on a given install date. A high value for session count on a given install date means that there are a lot of playthrough activity

MKTExpenses	This is the total amount of marketing expenses, in USD, spent to advertise the game. Given an install date, the marketing expense normally determines the cohort size. A high marketing expense means more advertising channels have been used to target more potential users to install the game.
ActiveUsers	This refers to the total amount of unique users who spent considerable time in the game given a certain date. This refers to the "stickiness" of the application.This is one of the attributes essential for determining a game's success.
ActiveUsersDay7	This is similar to the *ActiveUsers* variable but offset 7 days after the install date. This is the variable to be **predicted**.

Predicting Daily Active Users for Day 7

In reality, given a install date, and one would like to know how many daily active users would there be **7 days** after, the following variables will be used: Cohort Size, Day 1, CrashesANRDay1, DailyAverageRating, LevelPlayedEvents, LevelSuccessEvents, LevelFailedEvents, Sessions, MKTExpenses, and ActiveUsers.

Note that some variables like DailyAverageRating and CrashesANRDay1, only becomes available a day after. In a practical scenario, one could make predictions by Day 2 since it is assumed that all variables are readily available.