

## Analytics and Data Science @ Pocket Gems: Homework Instructions

### Overview

The majority of our users never generate any revenue (i.e., ‘convert’, in marketing parlance). We want to offer a promotion to some of these ‘non-payers’, in hopes of generating incremental revenue. Unfortunately, it’s not easy to identify true non-payers because many payers don’t convert until several weeks after installing the game. If you simply set a cutoff date and offer sales to everyone who hasn’t yet converted, you might find that revenue actually falls. This can occur if large numbers of not-yet-converted future payers take advantage of the sale instead of paying full price. If you’re not lucky, you might fail to convert a sufficient number of true non-payers, while ‘cannibalizing’ future payers by providing more value than they actually require.

The goal of this assignment is to devise a scheme to maximize incremental revenue. We’ll actually specify the promotion to be offered, namely, a two-for-one sale. **What we want you to do is to specify a target group for the promotion, as well as build a simple machine learning model in Python to predict a user’s probability of conversion.**

### Data

We’ve provided sample data for a group of users who installed during the first quarter of 2019. The data includes:

- User data, including user ID and install date
- Session history, including date and session number
- Purchase history, including date and amount
- Spending (and earning) history, including date, currency, and amount

‘Spending’ takes place when a user exchanges game currency (usually ‘gems’) for something of perceived value, such as a new outfit. Game currency can be earned during gameplay, but the bulk of it is obtained by making cash purchases, which are recorded in the ‘iaps’ (in-app-purchases) table.

The data itself is contained in four .csv files:

- users.csv: user\_id, install\_date, lang, country, hw\_ver, os\_ver
- sessions.csv: user\_id, ts, date, session\_num, last\_session\_termination\_type
- iaps.csv: user\_id, ts, date, prod\_name, prod\_type, rev
- spendevents.csv: user\_id, ts, date, story, chapter, spendtype, currency, amount

Note: Revenue is recorded in the ‘rev’ field and is measured in cents.

Additional Note: To fully understand how the spendevents table works, please filter to the user\_id '2062', and then order the results by ts. You'll see three types of spendtype:

- earnGemsCounter records where the user earns gems. In these records, the amount field is negative, indicating that gems are flowing out of our bank account.
- IAP records where the user purchases gems. In these records, the amount field is also negative, again indicating that gems are flowing out of our bank account.
  - In this case, the revenue is recorded in the iaps table and a corresponding entry is made in spendevents.
- PremiumChoice records where the user spends gems. In these records, the amount field is positive, indicating that gems are flowing out of the user's bank account.

### Assignment

1. Specify a target group of users for the promotion. Briefly justify your choice using the sample data. (Recommended time: 60 mins for exploratory data analysis, 30 mins for deciding on parameters of the target group.)
2. Build a simple machine learning model in Python to predict a user's probability of conversion. Please note: we're not looking for anything complicated here — we're more interested in how you build a model, which data you choose to train the model, and how you define your target variable. Please *do not* spend a lot of time obsessing over feature selection, feature engineering, or hyperparameter tuning. (Recommended time: 90 mins.)

Note:

- The promotion is a two-for-one offer. We will not be reducing prices. Instead, we will be providing twice as much value for the same price.
- There are no right answers to these questions. Please use your judgment and explain your reasoning clearly — that's what we're most interested in. The answers might be more subtle than you think. Additionally, please be brief — in our experience, the best answers to this assignment have been the most concise ones.
- **Important:** We expect this assignment to require approximately 3 hours of effort. Feel free to submit whatever supporting materials you'd like, in any form that suits you. **However, to be considered for the position, you *must* summarize everything on a PDF of 2 pages or less.** Also, please attach the code for the model separately, either as a .py file or as a Jupyter notebook (.ipynb, .pdf, or .html). Finally, be sure to include your name in the filename of every submission. Thanks, and good luck!