**Data Quality Issues:**

* A user makes an in-app purchase, but that purchase is not recorded in the spendevent.csv (51 users have this issue, i.e spendevents file failed to record iap and gems flow for 51 paying users). Exclude these users.
* I am interested in 100 but these 100 non-paying users don’t have any records in spendevents.csv. I don’t know if this is the error in database or because they didn’t make any actions (even though they are very active in the game). I excluded them in any process since most of important information is in the spendevents.csv for non-paying users.
* A lot of duplicates in session.csv, iaps.csv, and spendevent.csv

**Assumption**: I assume that we want to find potential groups at the current time in the dataset (i.e '2019-05-06'), so I will remove all the players who are not active in the last 7 days because there is a very high chance that they already churn).

Problem 1: Return a group of potential users

Chart, histogram, scatter chart

Description automatically generated

Insights: almost 90% of paying users made their first iap within the first 10 days.

Based on what we discuss so far, we will send promotion to people whose behavior are as follows:

* After removing all non-active players within the most 10 recent days, we left with 1733 non-paying users. Users who are not active within 10 days are most likely to churn.
* Given that they are still active within the past 10 days since the most recent date in the data (which is 2019-05-06). Since all the install\_date starts from 2019-03-01 to 2019-03-07, all active users are in their 50-60th days of playing the game. 60% of them has level less than 10 and low number of gems (<50). 80% of them have session\_num less than 250.
* Promotion group (for sample data): send promotion to users who are still active within the last 10 days i.e (from 2019-04-29 to 2019-05-06), have level/chapter less than 10 and number of gems less than 50 and max session\_num is greater than 100. There are 254 users whom satisfy the above condition.
* Reason: Consider all the paying users who made their first iap after 50 days in game, 90% of them has level lower than 10 and 80% of them has less than 50 gems. We are learning the pattern in paying users, and choose non-payers whose pattens are similar to paying users.
* In our sample data, all the active users within the most recent 10 days are in their 50th day of the game. In reality, we can target active users within the range of 7-40th day of their lifetime, then depends on the level and number of gems, we can send promotion.
* We shouldn’t send out the promotion within the first 6 days since that may help a portion of player to buy at the lower price than the price they are willing to pay for, but we may so do to keep a number of potential churn users to stay with us longer (there are 9k users quit the game after 6 days of installing)

Problem 2: Build a simple machine learning model to return the probability of user conversion.

As discussed in problem 1, we will remove all users who made a first iap within 7 days since installed date. Since those are people of “sure-things” group. We want to detect the pattern of “persuasive” group. We will also remove users who haven’t converted to a payer after 31 days since they are less likely to respond to our promotion.

Modelling: (assume current time is '2019-05-06'. Recall that all the install\_date starts from 2019-03-01 to 2019-03-07)

* In our sample data, all active people within most recent 7 days are already in their 50-60th day of the game. The chance they convert to paying users are not high. I couldn’t find any active users who are in their 7th-30 days of their game. And we don’t want to target non-active users since they already churn. Here is what I think is the best:

Train data (in our sample data):

* Take all the payers who made their first purchase after 7 days -🡪 label 1 (239 users). For each of these users, the look backs window is 7 days before the day they made their first iap (we want to learn their behavior of paying user before they make purchase). We will calculate features based on that 7 days.
* For those who are not active within the most recent 10-25 days 🡪 label 0 (we chose so because those users were playing the game for a reasonable amount of time (they were in the game in around 40 days and they haven’t made any iap and they churn). (total of 1386 users)

Test data:

* We will predict on those who are active within the last 7 days (all of them are in their 50-60th day of game) (total of 1067 users)

Alternative train data (general version):

* if we have more data, I think it is better to take all the non-paying users who haven’t converted after 60 days as label 0 as well, since after 60 days, it is a very small chance they will convert. Label 1 is from the paying users we have recorded. We will be targeting people in their 7 – 30th days of game and have low level and small amount of gems.

Model:

* Calculate features for each class
* Train a simple logistic model with class\_weights
* Analyzing model

Feature Engineering:

* We will be building a very basic model, just a few features relating to the number of sessions, what time in the day users play the game, changes in amount of gems, etc.