Cedric Herman
Mentor: Srdjan Santic
Career coach: Allison Matthews

Springboard
January 2018 cohort

Capstone #1: Data Wrangling

KKBOX (streaming music service provider) made available 3 files to get to know their customer base behavior. Transaction CSV file contains transaction history from January 1st, 2015 to February 28th, 2017. Information about plans subscription and prices are included. User log CSV file captures user's daily activity such number of songs played. And finally, member CSV file features demographic information on users who created an account on KKBOX.

1) Transactions

Based on roughly 21.5 Million transactions, more than 50% are completed using payment ID 41. Unfortunately, payment methods are represented by numbers only, so we don't know what it means. Intuitively, I would think this is a credit card automatic payment.

In this 2 years and 2 months period, membership automatic renewal is authorized 85% of the time. Looking at KKBOX website (thanks google translate), there is a discount when automatic renewal is active which can explain its popularity. Users are allowed to cancel at any time, this is recorded as a Boolean. An overwhelming majority of transactions do not show a cancellation (~95%). This is consistent with the fact that churn rate is very low (e.g. 6% in February 2017). Also, it happens that a user chooses to cancel its current plan to join another one.

Plan's duration in days shows a wide variety compared to KKBOX website where 30 days, 90 days, 180 days and 365 days are listed. However, KKBOX does offer bonus days which adds multiple combination and their special offers may have changed over this 2 years period. The most popular plan by far is 30-31 days followed by 7 days. It extends up to 450 days. Note this data has 0-day plans which seems erroneous, it represents 4% of all transactions. Taking a closer look at 0-day plans, all of them are also missing list prices which have a value of 0 (We will worry about list prices later). However, all other information seems to be legitimate thus we can calculate the number of days elapsed between expiration date and transaction date which should give us the plan duration by definition. Note that this can be done only when there is no active cancellation otherwise expiration date is not meaningful. Unfortunately, doing so produces many new values for

payment plan days, some of them occurs only once and some are even negative! The validity of this method is questionable. So, we will consider replacing 0-day by calculated values only if these already exist in our data to be safe. In order to tackle 0-day with active cancellation, we need to look at transactions grouped by users. If a user cancelled, it means he/she signed up for a plan first so we should see what plan it was from earlier transactions. TO BE CONTINUED

List price and actual amount paid in New Taiwan Dollar (NTD) are also available. List prices are rather discrete (0, 99, 100, 129, 149,150, 180…). Only 2% of transactions are greater than 180 NTD, these mostly corresponds to long term plans. In order to visualize list price and plan duration, we will first categorize plan duration using the following intervals:

| 0 - 7 | 8 - 29 | 30 - 89 | 90 - 179 | 180 - 364 | 365 - 450 |

TO BE CONTINUED