Music Recommendations: How UX research Improved KPIs

#Quantitative #Experimental design #Statistical analysis

Context

In the previous qualitative research, I gained a new insight to increase streaming counts, an important KPI of recommendation team. It was adding users' familiar songs at the top of the list. However, my suggestion was against data scientists' common view of "filter bubble".

This "filter bubble" issue in recommendation systems can cause user attrition when repeatedly suggesting the same popular items without variety, limiting discovery. A well-known approach to solve this issue is to increase the diversity in recommended items, which was exactly opposite to my insight.

Challenges

- Internal stakeholders were skeptical to adapt the UX insight into action.
- There was a chance that directly applying the insight might trigger a filter bubble.

Actions

My first strategy to convince data scientists was to admit their perspective first. Then, I tweaked the question to focus on diversity instead of familiarity: **"IF the diversity in recommended lists increases, would users listen to them more?"** They were convinced by this question and allowed me to do research on it.

To solve the problem, I chose a quantitative approach because the question was about "How much/many". I chose A/B testing to minimize any negative effect from a filter bubble.

Research process

Log data analysis: I analyzed log data to define user groups based on their behavioral tendencies, in particular whether a user was "more" and "less" explorative. I developed a formula to measure users' explorativeness.

A/B testing: I designed an A/B test to investigate the impact of diversity. First, I defined what "diversity" measures for the experimental design. As ML engineers ordered playlists by similarity scores from high to low ones based on users' streaming records, I adopted the system and defined that "highest" diversity meant "lowest" similarity score, and vice versa.

Hypotheses:

- Would users **listen to** more diverse songs over less diverse ones?
- Would the more explorative group listen to diverse songs more than the other group?

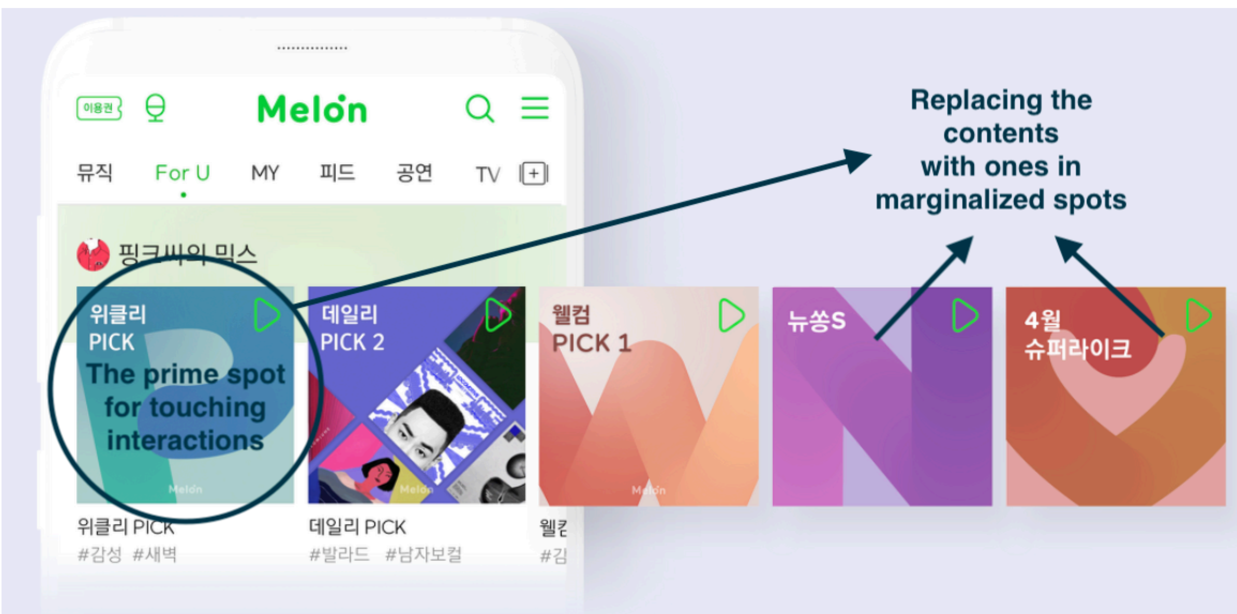
Variables:

- Independent: Level of diversity and user explorativeness
- Dependent: **Streaming count** and other minor KPIs - CTR and PV (page views)

2x2 Experimental design:

- More vs. less explorative user groups
- Test vs. control condition
    - Test: The first playlist is randomly replaced to one of other playlists.
    - Control: Songs in all playlists remain the same.

As the first playlist was the easiest to touch on the phone, I suggested to keep the playlists' cover visuals the same but to change the contents only, to avoid confounding in UI design.



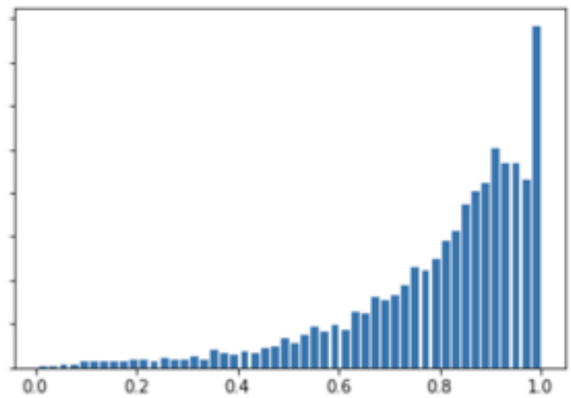
Results

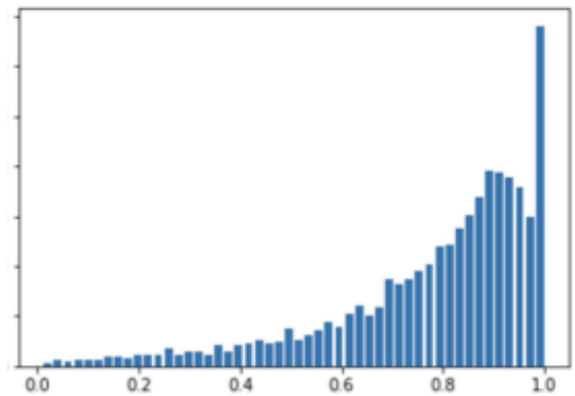
For log data analysis, I assumed that "explorative" users would prefer diversity than those who aren't. I devised a formula to measure users' explorativeness based on their choice:

the number of new songs / (the number of new songs + the number of already listened songs)

The closer the ratio gets to 1, the newer the songs that users chose are. For convenience, I used the mean to separate the "more" explorative group from the "less" explorative group.

I checked sampling distributions by extracting 7-day log data and applied a systematic sampling method, and confirmed there was no significant difference in distribution between two groups before running the experiment.





For A/B testing, the experiment lasted 7 days to collect user behavioral data. First visualized by Python Seaborn and then analyzed using ANOVA, no statistically significant difference was found in streaming count despite only a minor increase.

However, more explorative groups revealed significantly higher CTR and PV particularly in test conditions. No interaction effect was found.

These findings imply:
1. Simply providing highly diverse playlists doesn't affect users' streaming behavior.
2. Diversity in recommendations seems to attract explorative users, which emphasizes the importance of understanding users' behavioral patterns.

Combining the results with previous findings, I suggested that **providing familiarity at the first sight** might enhance users' reliability of the recommendation system, not causing a filter bubble yet. "Apparently, they seem to know the songs I would love."

Impact

- **Internal stakeholders were successfully convinced by the results and adopted the initial insight.** They learned increasing diversity may not be the solution for all cases.
- **The main KPI was increased by 6%** by adding a few familiar songs on top of the lists.
- **The team gained other insights that may lead to business impact:**
    - Personalizing diverse songs for more explorative users to increase retention

- Adjusting the level of familiarity depending on users' product lifecycle - provide familiar songs in the beginning of use, then increase diversity over time to prevent user attrition due to filter bubble