The Effect of Spotify Premium Service on User Engagement

Authors: Rachel Chen, Shirley Gui, Yuheng Ling, Jiayi Nie

Research Question and Background

Spotify, as one of the world's leading digital music, podcast, and video service providers, has acquired numerous premium subscribers over the years. As a premium subscriber, users can enjoy ad-free and high quality music together with the possibility of offline downloads. In this project, we are interested in whether premium can effectively increase user engagement, which can be represented by the session length in our dataset. Our hypothesis is that the premium service can enhance user experience, leading to longer listening hours per session, hence a higher session length in records. In order to test the hypothesis, we will utilize both Standardization and Inverse Propensity Weighting to estimate the treatment effect, and make comparisons between the two resulting average treatment effects.

Data Description

We are going to use *log_mini.csv* (session logs info) and *tf_mini.csv* (track's metadata and features) to perform our causal analysis. As for the log_mini dataset, it contains information about the session, specific actions when listening, and user type (whether premium or not). The tf_mini dataset lists the characteristics of each track, evaluating every track from 30 aspects like loudness, speechiness etc.

* The original dataset is over 50G, thus we are not able to handle it on our local computers, so we use the mini version instead.
* Datasets are downloaded from: https://www.aicrowd.com/challenges/spotify-sequential-skip-prediction-challenge/dataset_files

DAG

The most preliminary work before performing causal analysis and identifying suitable methods is to formulate the Directed Acyclic Graph (DAG), detailing the dependency relationships between features. We formulate the DAG below (Figure 1) from analyzing useful features provided by the two datasets together with our own experiences using Spotify and online resources. While some features are only dependent with one of either the treatment variable (*premium*) or the outcome variable (*session_length*), some others (*track_agg_score*, *us_popularity_estimate*) are confounding variables as these two influence both the treatment variable and the outcome variable.

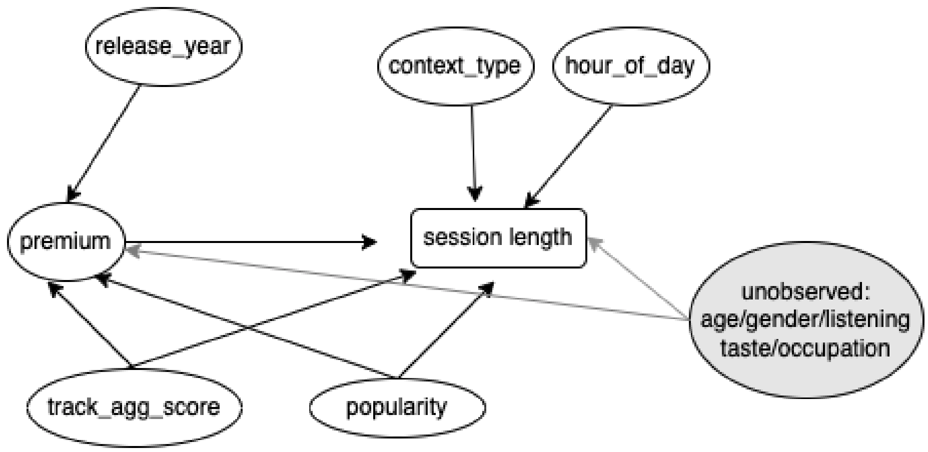

Figure 1: DAG

Data Cleaning and Exploratory Data Analysis

Our data cleaning and exploratory data analysis process is based on the DAG (shown in Figure 1). Firstly, we select relevant track characteristics columns from the tf_mini dataset, standardize those columns, and generate *track_agg_score* by taking the mean of each row (each track). Secondly, we combine the two datasets based on track_id using inner join and select only columns shown on the DAG that are going to be used in the causal analysis. Thirdly, we use one-hot encoding to encode the two categorical features (*context_type* and *hour_of_day*). The distribution of the two columns are shown below in Figure 2.

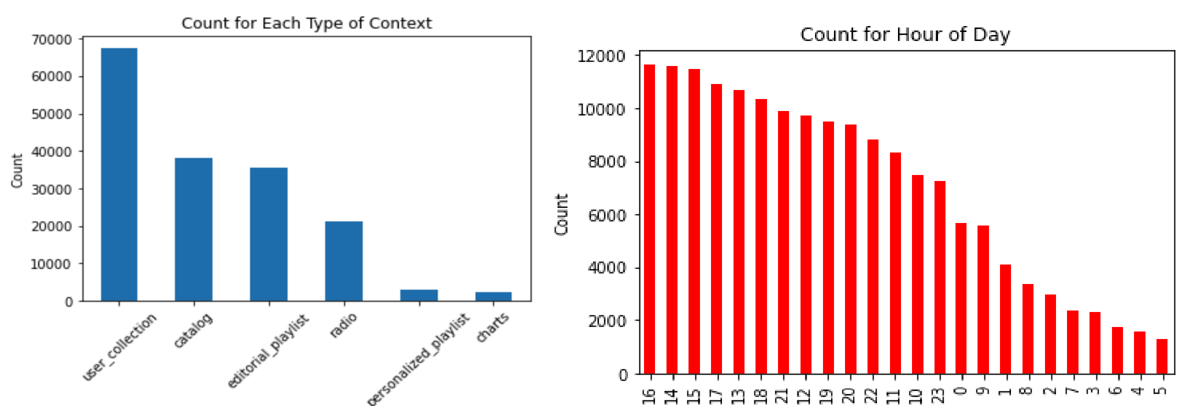

Figure 2: Barplots for *context_type* and *hour_of_day*

Fourthly, we standardize *us_popularity_estimate* and *track_agg_score* columns and change the premium column's representation to 0/1. Finally, we combine all the columns (after one-hot encoding or standardization) into a dataframe called *final_df* to be used for causal analysis. The distribution of the treatment variable and the outcome variable are shown below in Figure 3.

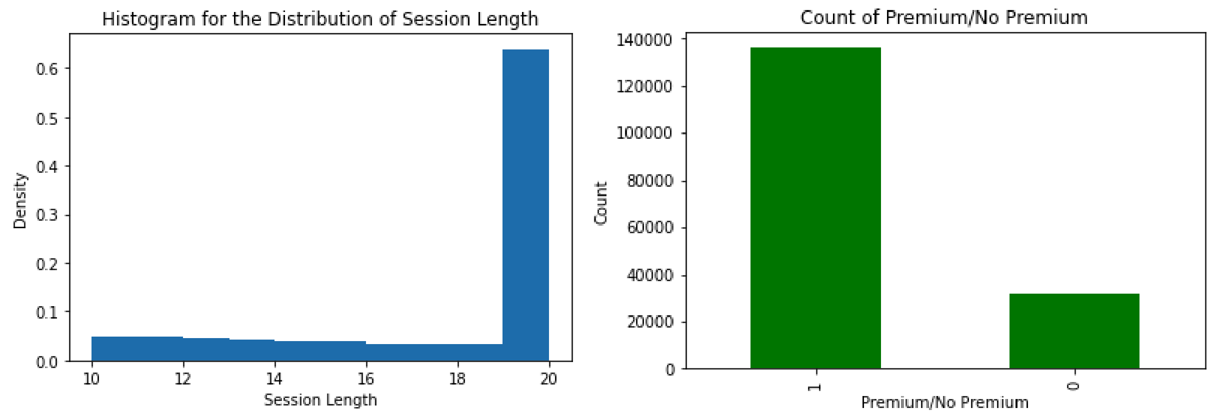

Figure 3: Histogram for the distribution of Session Length, Barplot for the count of Premium

<u>First Method: Standardization</u>

For the first method, we use standardization as a way to measure the causal effect of being a premium subscriber on *session_length*. In our case of analysis, conditional exchangeability can be achieved with observed confounding variables *track_agg_score* and *us_popularity_estimate* from the two datasets (shown in DAG).

We construct an Ordinary Least Squares model: '*session_length ~ premium + track_agg_score + us_popularity_estimate + context_type (in one-hot encoding format) + hour_of_day (in one-hot encoding format)*' to estimate $E[Y \mid X = x, L]$ given Y is *session_length*, X is *premium*, and L is the two confounding variables together with the variables that have direct impact on Y. After estimating the parameters of the OLS model, we set the value of *premium* to 0 and 1 to form two new dataframes. The difference of average predicted outcome values of each of the two dataframes estimates the ATE.

The final ATE from performing standardization results to be around 0.7362. This result shows that being a premium subscriber has a strong positive causal effect on session length. That is to say, premium service can enhance user experience, leading to longer listening hours per session.

<u>Second Method: IP Weighting</u>

Due to the existence of the confounding variable, exchangeability does not hold. However, if we conditioned on the measured confounding variables, i.e. *us_popularity_estimate* and *track_agg_score*, then conditional exchangeability holds. In other words, when conditioned on these two variables, the session length under the potential premium subscriber among the treated group equals the session length under the potential premium subscriber among the untreated group.

The original population satisfied the conditional exchangeability, so we could use IP weighting to eliminate the bias. The idea is that popularity and score of a specific track may have an impact on user's preference in becoming a premium user. The original population is not quite representative of the broader population, so we create a pseudo-population. We use the propensity score, which is known as:

$$px = P(premium \mid us\_popularity\_estimate,\ track\_agg\_score).$$

We weights treated groups by $1/px$ and controlled group by $1/(1 - px)$. In our analysis, we used logistic regression of "*premium ~ track_agg_score + us_popularity_estimate"* to get the propensity score. After the weighting, we used WLS regression to get the ATE, which is around 0.7443. The result refines our assumption that the premium subscribers have a positive effect on the session length.

Conclusions and Potential Improvements

Based on the results from Standardization and IP weighting, the premium has a significant positive effect on the session length. Compared to the non-premium account, premium accounts will increase the session length by roughly 0.74, with a 95% confidence interval of [0.710, 0.778] (from IP weighting WLS summary table), meaning that we are 95% confident that the average treatment effect of premium will lie between 0.71 to 0.778. Therefore, we can arrive at a preliminary inference that premium subscription is efficient in terms of increasing users' engagement.

However, due to the nature of our datasets, additional improvements and considerations should be conducted in order to further confirm our results:

1. **Feature Selection:** We may consult experts in psychology or the musical industry to include more features that may potentially affect treatment or outcome variables. This will lead to a different causal analysis setup and more DAGs.
2. **Class Imbalance:** Our data is highly imbalanced, with ~130000 non-premium accounts and ~30000 premium. We may need to dig more to see if the existence of imbalance would trigger bias when performing causal analysis.
3. **Unobserved Confounding Variables:** There are many unobserved confounding variables affecting both the outcome variable and treatment(as shown in the DAG), our results will be more robust if we could acquire more information on users' data.

Link to code: (log in with lionmail)
https://colab.research.google.com/drive/1YPD5FmdH-0oL4eUt11b1AMlmyqkfzN9t?usp=sharing