Journey App: Notification Time Data Analysis - Deliverable 1

Group Member: Lingyan Jiang, Ruoqi Shi, Jiaqi Zhao

Data Collection:

Raw data was given by the client at first. Our team has supplemented the data many times and asked for new data. At the same time, the data is supplemented with the A/B test, which enables better data analysis. In the future, we can actually run our results in the application to verify our data analysis.

Data Preliminary Analysis:

1. Understanding the data:
    a. For the definition of data, since each company has its own idea of formulating data titles, the meaning of data may be completely different. Therefore, we ask our customers in detail.
    b. We initially clarify the meaning of each data with the client:
        i. Info Customer ID - The ID number of the users' company, there are about 20 company IDs in total.
        ii. User Id - Each user's Id on behalf of which notifications are sent.
        iii. User-Created At - The first date of hire of the user to whom notification is sent.
        iv. Activation Date - Date and time when a notification is sent.
        v. Activity Date - Date and time when a notification is opened by the receiving user.
        vi. Name - The title of the notification content.
        vii. Content ID - the ID of the notification content name.
        viii. Content Type - The type of notification content.
        ix. Journey Name - The journey that the content of the notification comes from.
        x. Action - Percentage of content read by the end user.
        xi. Device - The device used by the user to read the notification. Due to the current universality and convenience of mobile phones, the focus is on mobile phone users.
        xii. Channel - The channel that users click on the notification.
        xiii. Session Id: When a notification is sent to the user, a session is locked in the system.

Data Preprocessing:

1. Converted all the categorical data using strings into numbers
    a. Columns included: Content Type, Device, Channel

2. Replaced all the NaN values by finding the MAX of the value of the categorized variable and replacing NaN values with the MAX value.

Data Analysis:

1. The questions our group have answered:
    a. What parameters/columns are playing roles in the action (notification is opened/ignored/completely read/not read) performed by the end user?
    b. Based on these parameters can we identify the optimal time for sending a notification to every existing user?
    c. What is each user’s most frequent time to open?
2. How did we analyze the data to answer those questions:
    a. Created a new data frame
        i. Our group created the new data frame with columns of User Id, Activation Date, Activity Date, Action, and User Created At.
        ii. We implemented a new column time lag which is calculated by the difference between Activity date and Activation Date.
        iii. We filled the missing data with different methods. For the missing content name, we used maximum count data to fill the missing content.
    b. Created a dictionary
        i. The key is the User ID; the value is the most frequent hour time when each user opens the notification
        ii. For each user, we store the most frequent hour time when they open the notification in this dictionary. This gives us the recognition of the data and the habit of each user.
3. What are some potential problems in this analysis?
    a. Since there is a possibility that the user will only open the notification but not finish the content, we should also take into consideration the “Action” column, which indicates how many percent of content did the user finish.
4. Next questions need to be answered:
    a. Can these insights be generalized to future coming users?
    b. What algorithms will we use for predicting future users?
5. Following steps to improve:
    a. Use the “Action” column to find the time when each user has the highest completion rate of the content sent by the notification.
    b. Considering the A/B test in the dataset, and may need to drop some data of one user based on the A/B test result.
    c. There are sometimes big time lags between the notification sending time and user reading time, we need to consider the time lag that influences the notification completion rate.