Data Scientist Exercise

Guideline

The goal of this exercise is to highlight your technical understanding of churn prediction model and how it can benefit our product management and/or online marketing team.

We do not expect you to come with a complete solution, but we would like to see your creativity, expertise and approach to challenges.

We suggest you to:

- Read and take your time to design a solution
- Choose how to present (mockup, wireframes, PPT, code, anything you like)
- Do not worry, the scope is to trigger a conversation, not to judge

Exercise

Bonial mobile users view a brochure from a retailer published on our platforms. Each time a brochure is viewed, we capture several data associated to the user, i.e. how the user engages with the content.

Given the datasets below, predict which user is still active in July based on the behavior of April until June. How would you approach that problem? What other data points would increase the accuracy of the algorithm? Which features did you select and why?

Description

- Data extracted correspond to all installs in April & May 2017
- 20,000 users in total, with corresponding app starts and brochure views
- Tab separated ASCII files

Dataset Files

- Installs.txt
    - A sample set of installations from April and May 2017
    - Productid – different native apps that we have on the market
    - userId – internal identifier
    - model – smartphone device name
    - campaignId – indicates how the user installed the app
- brochure views.txt
    - all brochure views of the users between April and June 2017
    - dateCreated – date of the brochure view event
    - view_duration & page_turn_count show the level of engagement with the content
- brochure views july.txt
    - same structure as brochure views.txt
    - all brochure views in July 2017
- app starts.txt
    - date and userId of the application start
- app starts july.txt
    - same structure as app starts.txt
    - all app starts in July 2017