Exploration of Celtra's Platform Usage Data

Urša Zrimšek

uz2273@student.uni-lj.si, 63200441

Introduction

Celtra is a leading company in the area of automated creative campaigns for digital advertising.

In this report we will explore and visualize two datasets, provided by Celtra company. One of them contains platform usage data and the other shows us advertisement traffic on Celtra servers.

First we will look at the structure and size of the data and then we will visualize it, to better understand the usage of the platform and sessions that the users have on it.

There is some more information about data and some visualizations and calculations to prove findings from this report in Jupyter notebook `data_exploration.ipynb`.

Data representation

Celtra company provided an anonymized and sampled dataset on their platform usage data. Let's overview the attributes provided in both datasets.

Platform usage data

Platform usage activity is shown in a data frame with 622078 rows, each represented with 6 columns:

- `ACCOUNT`: unique ID for each company,
- `USER`: unique ID for each company employer,
- `SESSION`: unique ID of session using the platform,
- `ACTIVITYLOCATION`: part of the platform used,
- `ACTIVITY`: coarse grouping of `ACTIVITYLOCATION`,
- `TIMESTAMP`: time of usage.

If we look at unique values in each column of the data frame, we can see that in the 90 days of gathering this data, from 1.7.2020 to 29.9.2020, 11 different companies with 266 users were using Celtra platform, and they altogether participated in 62693 sessions. They used 48 different activity locations, and 10 different activities. Their names can be seen in the legend of Figure 1.

Sessions data

More visualizations will be connected to the second, bigger, data frame, represented with 4823186 rows and 15 columns. Some of the columns are explained here:

- `UTCDATE`: UTC date of gathered data,
- `ACCOUNTID`: unique ID for each company,
- `CAMPAIGNID`: unique ID for each campaign,
- `CREATIVEID`: unique ID for each creative,
- `PLATFORM`: platform where ads are shown,
- `SDK`: software environment on device where ads shown.

After each ID column for account, campaign and creative we have a column with creation date of that account/campaign/creative. Last 6 columns of the data frame represent numeric data about success of the sessions. First of them tells us the number of requests to the server from the users' devices, next one the number of successfully loaded ads and next the number of successfully shown ads. Then we have the number of ads that the internet users interacted with and after that the time (seconds) that the ads were visible to users. The last column tells us the number of ads that were attempted to be loaded, but it is always the same as the number of requested sessions, so we didn't use that column anymore in the further analysis.

If we count the unique values in the columns, we see that we have data for two years, 2018 and 2019, and in those years sessions were requested by 317 different companies, during 15937 campaigns with 62081 creatives altogether. The ads were showing on 6 different platforms: Android, DesktopPlatform, IOS, WindowsPhone, BlackBerry, WebOS, and in 8 different software environments: MobileWeb, MRAID, VPAID, VAST, AppleNews, SafeFrame, AMP, Pandora.

Exploration and visualization

In this section we will explore and visualize the data, to have a better understanding of it.

Platform usage

When visualizing platform usage, the most interesting was the usage by hour of the day, shown on Figure 1. Usage by the day of the week was expected. The platform is used less on Friday, and is almost not used on Saturdays and Sundays. Interesting thing is, that Sunday is the only day when the most used activity is reviewing, other days it's using Ad Builder. We have data for only three months, so we can't conclude anything about usage by the day of the month.

We can see which activities are most used on the Figure 1, the exact order is the order used in the legend. Here we should add that all but the last three are used by all accounts, *account/user managing* is used by 10, but the last two are used only by 4 and 2 accounts, the last one even only 14 times. In one session there was at maximum 5 activities performed, at 10 activity locations.

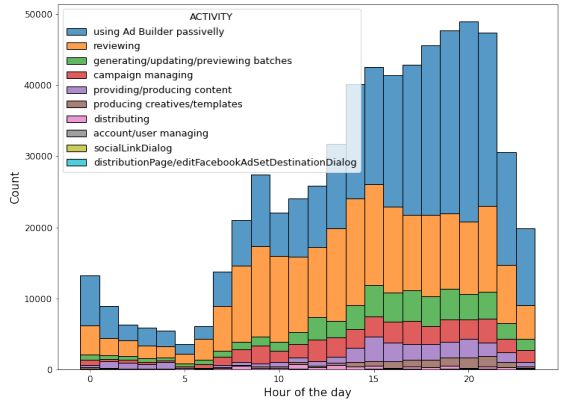

Figure 1. Usage of the platform by hour of the day. Here we can see the number of actions users made at certain hour of the day. The actions are separated by activity they performed. Activities listed in the legend are ordered by the summed number of times performing that activity. We can conclude that by far the most used activities are using Ad Builder (raising towards the end of the day) and reviewing (more constantly from 8.00 to 22.00), and that the busiest time is from 14.00 to 22.00. This shows that the majority of the platform's users live west from the UTC time zone.

Looking at account data, we can see that there is one really big account, that had twice as many sessions as the next one and performed more than 3 times more actions. The average account has 25 users, has used the platform 56553 times in 5699 sessions, in 33.6 different activity locations and has performed 8.5 different activities. But the standard deviations here are very big (except for activity 0.9 and activity location 4.8), so this doesn't tell us much about how an account should look like.

If we group the data by users, we can see that the most active user is *(anonymous)*. This is any user that used the platform without logging in. We can't make any assumptions about an average user, since some users could always use this option when accessing certain parts of the platform, that are available without logging in. Anonymous user was used by all 11 accounts and was used for 7 activity locations and 5 activities. When using those activity locations, users still logged themselves in almost 77% of times.

Sessions

When visualizing sessions we first compared the numbers for years 2018 and 2019.

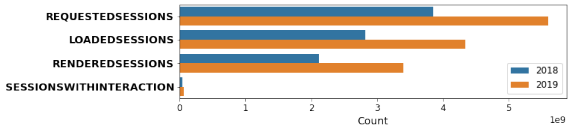

Figure 2. Comparison of sessions in 2018 and 2019. We can see that there was more session requested (and also loaded, rendered and interacted with) in year 2019, but when we calculate the percentage of successful sessions (sessions with interaction / requested sessions), it is smaller for 2019 (1.16%) compared to 2018 (1.2%).

On Figure 3 we can see how the number of requested sessions changes during the year. If we make the same plot for sessions with interaction and viewable time, it tells us only that when there is more requested sessions, there is also more interacted with sessions. So we rather looked at the ratio between interacted with sessions or viewable time and requested sessions. For both the ratio is a lot smaller where January and October spikes are. The viewable time ratio is the biggest on 8.8. and sessions with interaction ratio spikes just before Christmas and in the end of January, March and April. Average success (percentage of requested sessions that were interacted with) by the weekday is the smallest (1.13%) on Monday and highest (1.25%) on Sunday.

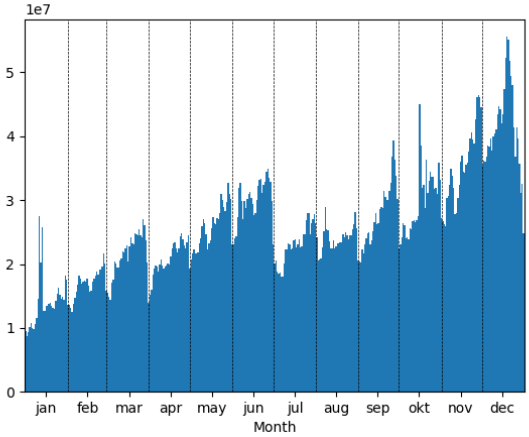

Figure 3. Requested sessions during the year. The number of requested sessions is summed for years 2018 and 2019. We can see that the number is increasing towards the end of each month, end of each quarter and towards the end of the year. The only month when the number substantially decreases in the end is December, with the peak of the whole year just before Christmas, on 22nd and 23rd. There are also big spikes in the beginning of January and in October.

Campaigns

Now we will focus on separate campaigns - how they look like and their success. We measured success as the percentage of requested sessions, that were interacted with. We didn't

take into account the viewable time, because someone could be visiting a site with this add without being interested in it.

First we wanted to know how long are our campaigns. Because some of the campaigns started before the data acquisition and we can't know how successful they were at the start, we limited this observations to those that started in 2018 or later - 12746 campaigns. We can see their lengths on Figure 4.

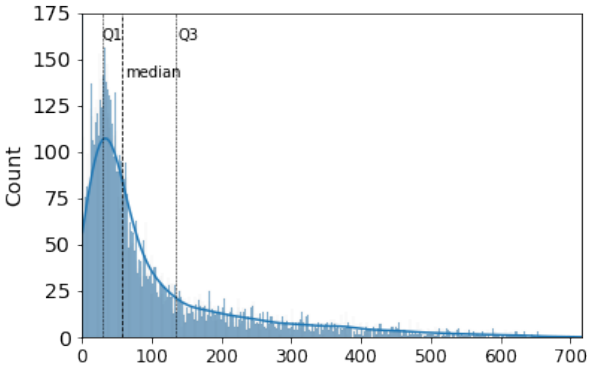

Figure 4. Length of campaign in days. Histogram of the lengths of campaigns started in 2018 or later. The shortest campaign lasted only for one day and the longest for 716 days. We added kernel density estimation to estimate the distribution of lengths, and vertical lines at quartile values - Q1 (29), median (57) and Q3 (134).

Now we were interested how the success of the campaign changes with respect to the number of days from the start of campaign. We can see that on Figure 5. On this figure we didn't draw any regression line, because we didn't want to lose any properties already visible without approximation.

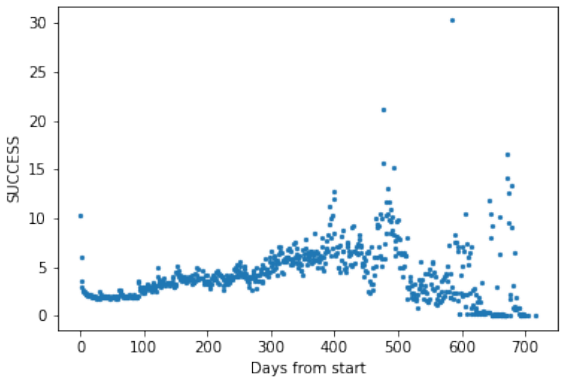

Figure 5. Average success of campaigns on certain day from their start. On the first day of a new campaign the average percentage of interacted with requests is over 10%, then the next few days success drops very fast, after 100th day it starts rising again. After 400th day values start jumping, because we have less than 600 campaigns over 400 days long, only 255 over 500 and 75 over 600.

The last thing we will analyze here is the success on different platforms. There is a big difference between the number of campaigns that were advertised on each platform and also their success. There was 14872 campaigns advertised on Desktop, with average success of 7.6%, 12963 on Android (3.8%), 13874 on IOS (4.5%), 7127 on WindowsPhone (1.0%), 4674 on BlackBerry (0.75%) and only 696 on WebOS (0.17%).

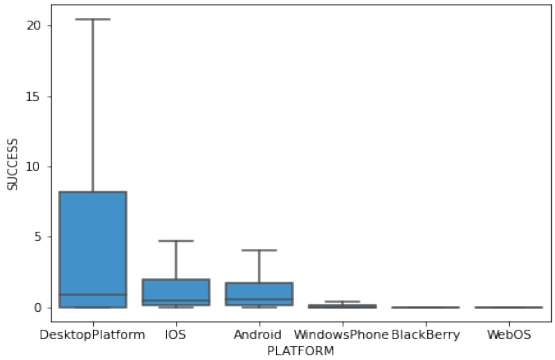

Figure 6. Success on different platforms. We calculated success of each campaign on each platform where it was advertised. The platforms are sorted by their average success. We skipped the outliers, because of visibility. The furthest outliers of the first 5 boxes have success of 100%, the last one at about 70%. The plot with the outliers can be seen in the notebook.

Discussion

When analyzing Figure 3, we need to be aware that we can't know which campaigns were still active in 2020, so there can be some error - campaigns could be longer than we are showing on Figure 3. It would be useful if we would also have data for the closing date of the campaign. If we would have the data, we could also analyze at what time of day people interact most with the ads, so the users would know when to push their creatives.

We know that the December spike in Figure 3 corresponds with Christmas advertising, but we should also find out why are there spikes in January and October. The reason for January spike could be the Orthodox Christmas. There could be a reason for October spike, or it could also just be an abnormality, since we only have data for two years and can't make definitive conclusions.

In sessions data we didn't pay attention to the number of loaded and rendered sessions, because Celtra or the company that is using the platform is only choosing the number of requested sessions, and they want to maximize the interaction.

Further we should explore and find out what makes some campaigns more successful than the others. There are some campaigns that only have a couple of requests and 100% success. Those could be test campaigns, and it would be better to leave them out in any predictive models.