Databox Dataset Insights

Matej Miočić[1]

Abstract

In this report, our goal was to provide insights into the data provided by Databox. We focused on paying companies and what makes them become one. With gained analysis we then focused on companies' sessions. We show a pattern based on 3 milestones of a company – date of sign-up, date of becoming paying and date of cancelling payment.

Keywords
Databox, Sessions, Becoming paying

[1] *mm9520@fri.uni-lj.si, 63180206*

Introduction

In Project 2, we were tasked to explore and provide insights into the data provided by Databox. Our goal was to analyse the dataset and describe it.

Description of datasets

We were tasked to explore data based on 2 datasets. First dataset includes general data about a subset of customers. It has 101,698 rows with 21 attributes as rows (see Table 1). Any unknown values were marked as 'undefined'.

Table 1. Table of attributes for the first dataset.

Attribute	Description
distinct_id	Customer id
space_id	Unique customer id
country	Country of the company
is_agency	Company is an agency
company_created	When company was created
became_paying	When company became paying
became_pql	When was categorized as PQL
cancelled	When was account cancelled
had_trial	Company used the trial option
trial_features	List of trial features used
is_activated	Connected a data source
level_achieved	Onboarding level achieved
has_mobile	Logged into mobile app
sessions	Sessions count
databoards	Databoards count
cloud_sources	Cloud sources count
metrics	Number of Metrics added
users_in_space	Number of users added to space
scheduled_snapshots	Number of scheduled snapshots
custom_queries	Number of custom queries
scripts	Number of scripts used

Second dataset combines event data for 8 different events for 2 years. It has 7,948,592 rows with 4 attributes (see Table 2).

Table 2. Table of attributes for the first dataset.

Attribute	Description
Date	Date of event in format yyyy-mm-dd
Event	Type of the event
Space	Id of the customer triggering event
Event_count	Number of events on the specific date

Events include: *CalculationCreated, MetricAdded, NewDatasourcesAdded, QueryCreated, Sessions, ShareUrl, SnapshotShared and UsersAdded*. It is worth mentioning that some events had begun logging at different dates. Most events started logging on the same week; 25.11.2019, but the first logging of sessions started on 20.04.2020. Both datasets include a few mistakes. For example in the second dataset 4 events are named "100023", which could indicate even more errors that we did not discover.

Analysis of paying companies

We decided to focus more on the analysis of paying companies. At first we looked at how trial features influence companies on becoming paying (see Figure 1). We found out that the most common used trial feature is *Query Builder* by far. We compared features based on how many companies become paying after they use this feature. We found out that *Query Builder* is at the bottom with only 5% companies that use it become paying. This could mean that the companies are not satisfied with this feature or they already did everything they need in the trial, so they decided not to continue. On the other hand, companies that use trial feature *More Data Sources* are

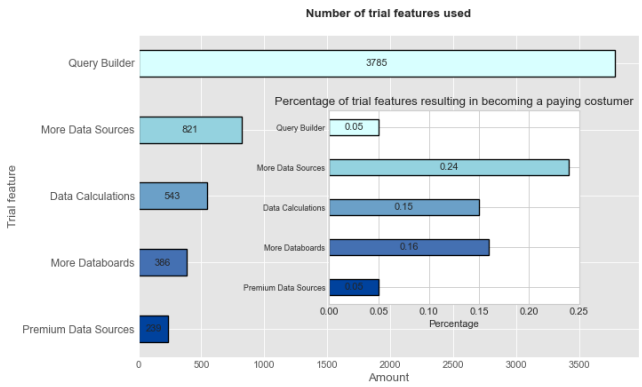

Figure 1. Comparison between how many companies use individual trial feature and what percantage leads company to becoming paying shows us that the most popular feature is not the most profitable one. It is one of the least profitable ones.

the most likely at becoming paying, with 24% users that use this feature become paying. When it comes to more data, we deduce that companies want more of it when they get a taste of it.

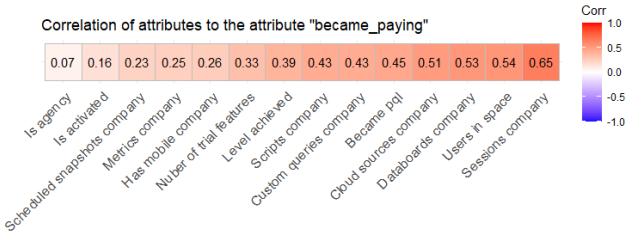

Figure 2. Correlation heatmap of attributes shows us what attributes are connected to companies becoming paying.

Next, we compared which attributes correlate more to the attribute *became_paying* (see Figure 2). Since we are not interested in specific dates when companies reached a milestone. We changed all dates into True if they exist and False if the value was 'undefined'. We also changed *trial_features* to *Number of trial features* which now shows how many features company used. From the plot we can observe that number of trial features is not as correlated compared to other attributes. The most correlated attribute is *sessions*, which tells us how many sessions users had. What we don't know yet is if number of session rises because company became paying or the other way around. To answer this question we continued our analysis.

Session analysis

To gain more insight from sessions we decided to compare mean values of number of session of each company based on a milestone. Most crucial milestone are; the week that company started using Databox, the week that company became paying and the week that company canceled payment. Since sessions in dataset are only available per week we concluded that the value corresponds to the week that follows. We decided to show 8 weeks period around a milestone. First we show mean of sessions for 8 weeks after a company started using Databox (see Figure 3).

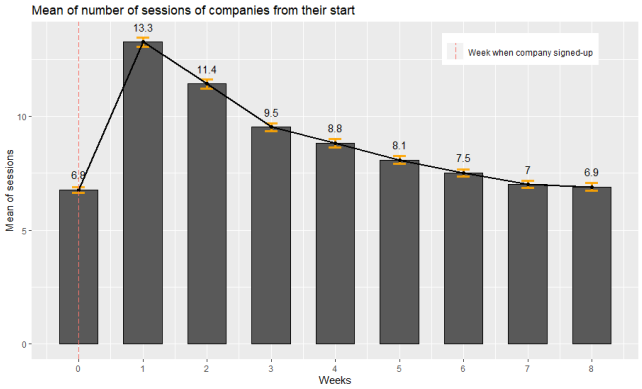

Figure 3. Means of number of sessions after the company signed-up with standard error, show us that companies start losing interest after time.

We annotate week 0 as the week when company started using Databox. Because a company can start using the site at the end of the week, we observe a low mean of sessions. That quickly changes the following week as mean rises and reaches its peak in the shown timeline. After that, mean expectedly starts to slowly fall. This can be explained by users being really engaged at the start and then slowly lose interest.

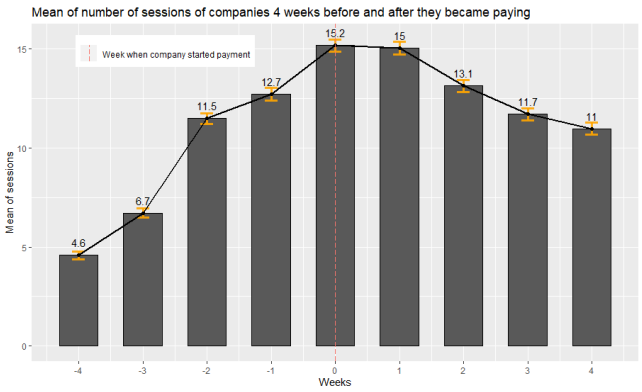

Figure 4. Means of number of sessions around company became paying with standard error, show us that on average companies reach their peak of number of sessions on the week they become paying.

For the next milestone – company becoming paying, we show a 4 week period before and after (see Figure 4). We annotate weeks from range -4 to 4, emphasizing that these weeks happened before or after the milestone. Surpisingly mean rises up until company becomes paying. After that it starts to decay. One would think number of sessions would

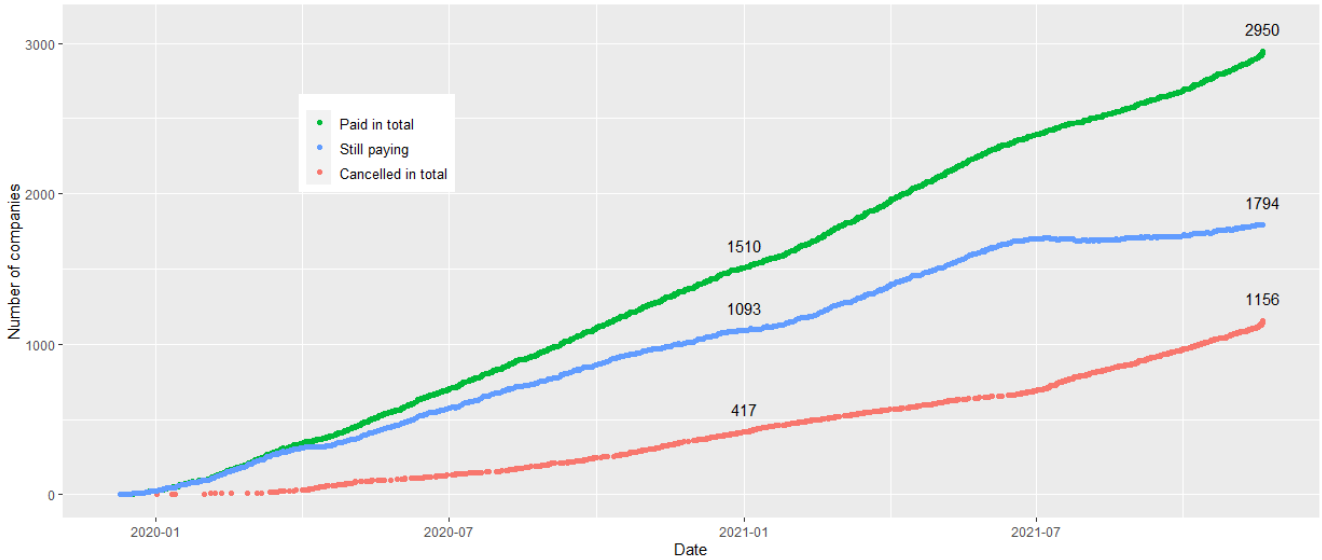

Figure 5. Number of companies that paid in total, are still paying at the time and cancelled in total show us Databox's progression in terms of paying users. It is seen that number of paying companies in total is almost linear. Number of companies that cancelled payment rises with paying companies. Hence the number of paying companies at specific time starts to stabalize.

rise after a company starts using premium features. It might be the case that companies do spend more time using Databox, but reset their sessions less often.

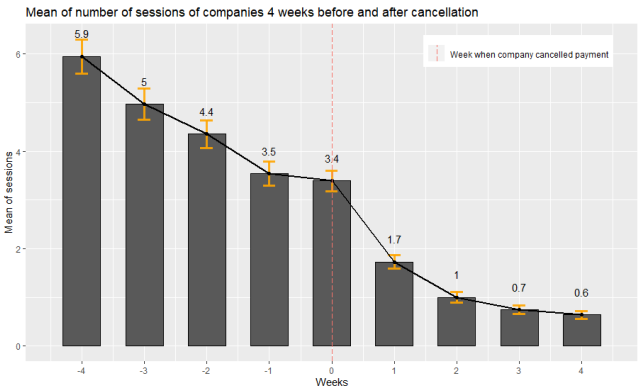

Figure 6. Means of number of sessions around company cancelling their payment with standard error, show us that companies lose interest before cancellation. After that they stop using the site almost completely.

Finally we show mean of session around the third milestone – company cancelling payment (see Figure 6). As in previous chart, we show a 4 week period before and after the milestone. We use negative numbers to show weeks happened before the milestone. In this plot we observe a decay in mean of sessions before cancellation. After the cancelattion, means drop to even less than 1 session per week.

For the final touch we looked into how Dataset is progressing in terms of paying costumers (see Figure 5). We have shown how many companies became paying in total, how many cancelled in total and how many are still paying sorted by date. We can see that with more paying companies number of cancelling companies increases. Number of companies paying at specific date eventually stabilizes in July, 2021.

Discussion

We analysed 2 datasets provided by Databox. With gained insights we can help the company understand data better and prepare for easier modelling for Project 3. We focused on paying companies and sessions. We found out which trial attributes lead to company becoming paying the most. Next, we found out that by looking at sessions we can see a nice pattern when looking at company milestones (date of sign-up, becoming paying, cancelling payment). For the final touch we have shown how Databox is performing when looking at paying companies throughout 2 years of data they have sent us. Since the 2 datasets are so large, there is a lot more we can explore.