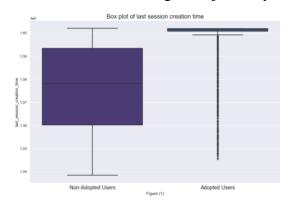# Report Summary

In this report, I will briefly introduce the approaches of finding the most important factors for predicting the user adoption and which factors they are.
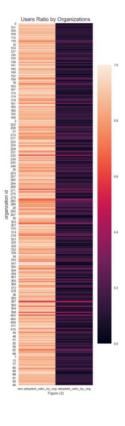
Before deep diving into the data, I did some basis data cleaning to guarantee highest quality information in the ultimate decision-making. There were 3177 and 5583 missing values in last_session_creation_time and invited_by_user_id respectively. I replaced last_session_creation_time's missing values with its mean, and considered users with missing invited_value_by_user_id were not invited by anyone, so I filled 'no' for the missing values. Subsequently, I converted all variables in terms of id from integer into string and variables in terms of date from string into datetime.

Next, identified which users are adopted in engagement data set base on users' login activities and inner joined the result and users data set by user id.
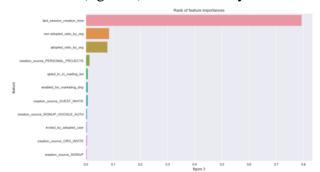
Through data visualization, I found that last session creation time and organization a user belongs are the best factor to distinct adopted users from non adopted users. In figure (1), the distribution of last session creation time for Non-adopted users is evenly distributed while the one for adopted users is extremely skewed and with much greater median, so last session creation time would be a great predictor. I was also interested in whether some organization have higher adoption rate than the others. The nature of an organization dictates a user's job content. Some job may require higher productivity, organizational and management abilities. Then users holing these jobs may be more likely to continue using the software. In figure (2), the heat map shows non-adoption rate and adoption rate by organization. The lighter color of a strip on the right side, the higher adoption rate for the corresponding organization. As we can see, there are some variations in the adoption rate over different organization.

Users Ratio by Organizations

Box plot of last session creation time

Figure (1)

Figure (2)

At the end, I also leveraged Random Forest Classifier to rank the importance of the factor (figure 3). The results by the random forest classifier is coherent with the exploration

Rank of feature importances

figure 3

above. Last session creation time is the most important feature deciding user adoption, followed by the organization a user belonging to. Besides, the creation source may matters the adoption. Users creating their accounts through personal project are less likely to continue using the software for long term, according to our sample data.

For further exploration, we can analyze the common characteristics of those organization with high adoption rate, and business can use these characteristics to find out more potential customers or users.