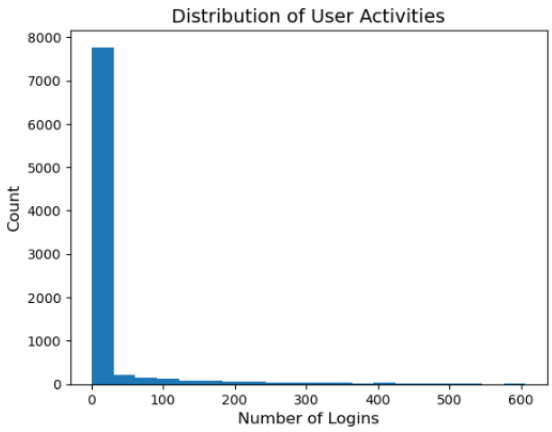
Relax Adopted User Challenge:

Problem Statement

Relax is keen on discerning the primary drivers that forecast future user adoption. Predicting user behavior holds paramount significance as it will shape Relax's forthcoming customer service and marketing initiatives. For this challenge, adopted users, who have logged into their product on three separate days in at least one seven-day period, need to be properly labeled according to user engagement logs. Then the primary features, based on user information, which drive user adoption need to be identified for future forecast.

User Engagement Distribution

Logins for each user show screwed distribution. The majority of the users only have 1 login.

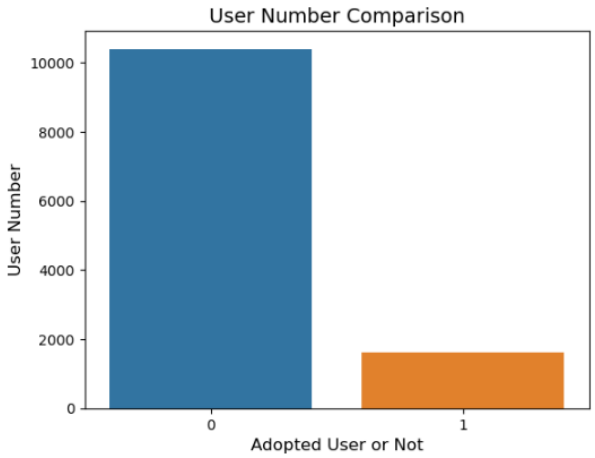



```
df_label.groupby('user_id')['date'].count().value_counts()
```

```
1      6235
2       340
3       241
4       181
5       110
      ...
332       1
86        1
395       1
509       1
255       1
Name: date, Length: 402, dtype: int64
```

Adopted User Labeling

There are 1,602 adopted users out of a total 12,000 users.



Data Wrangling & Feature Engineering

Missing values:

- `last_session_creation_time` has 3177 missing values; filled by `creation_time` assuming no logins for those with NaN `last_session_creation_time`.
- `invited_by_user_id` has 5583 missing values; filled by 0 assuming not invited by any existing user for those with NaN `invited_by_user_id`.

Feature engineering:

- A binary indicator if the registered email is in the top 6 domains or not.
- A binary indicator if a user is invited by an existing user or not.
- A numerical feature for user active time in seconds, which is the time between 'last_session_creation_time' and 'creation_time'.

Categorical values:

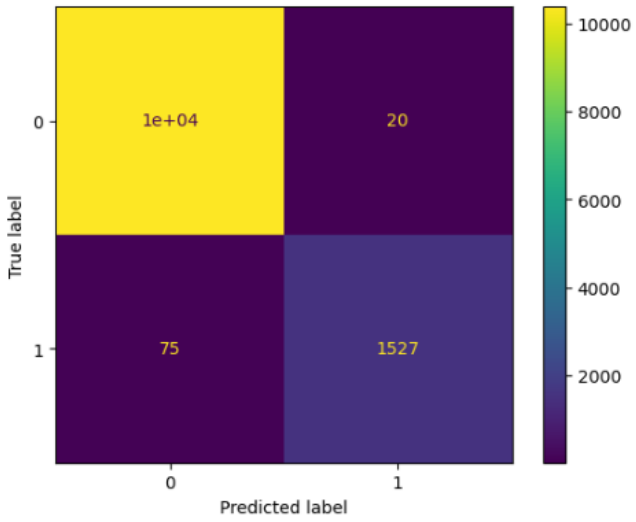
- Ordinary encoding for `creation_source`, which is a categorical feature

After data wrangling and feature engineering, the following features are finally used for predicting adopted users.

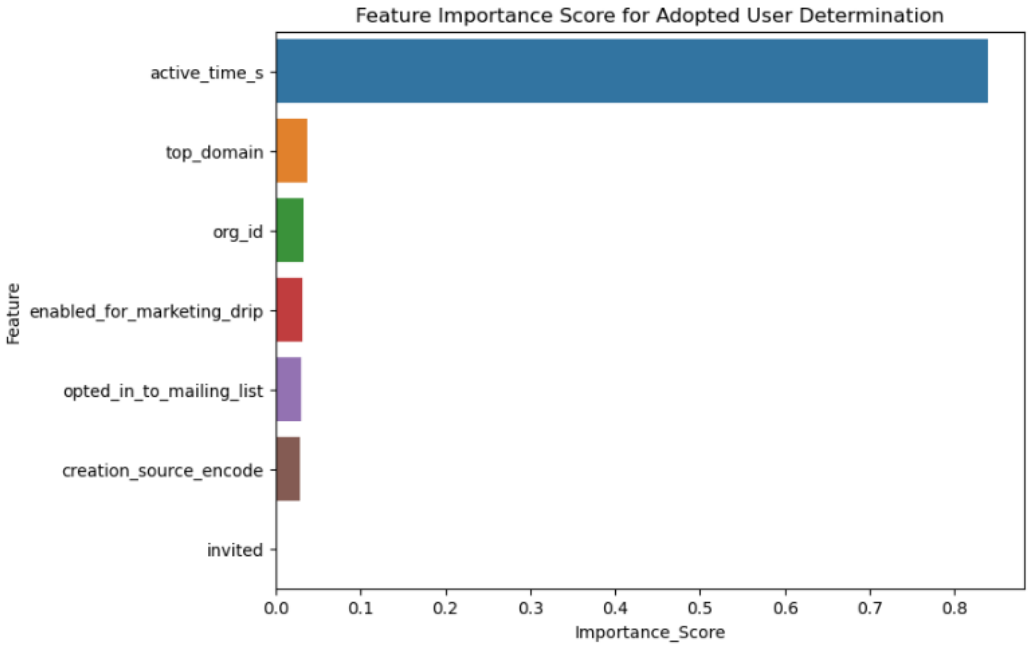
	opted_in_to_mailing_list	enabled_for_marketing_drip	org_id	top_domain	active_time_s	creation_source_encode	invited
0	1	0	11	1	0.0	0	1
1	0	0	1	1	11750400.0	1	1
2	0	0	94	1	0.0	1	1
3	0	0	1	1	86400.0	0	1
4	0	0	193	1	432000.0	0	1

Modeling and Feature Importance

The XGBoost model is used for this binary classification task. Here is the model performance.



The model can provide feature importance as well. Here is the feature importance comparison. The active time has the highest importance score.



Summary

The most important feature driving the forecast of adopted users is user active time. Adopted users tend to have a much longer active time, time between their last login and when they created the account.