Relax Adopted User Challenge

Fei Wang

Problem

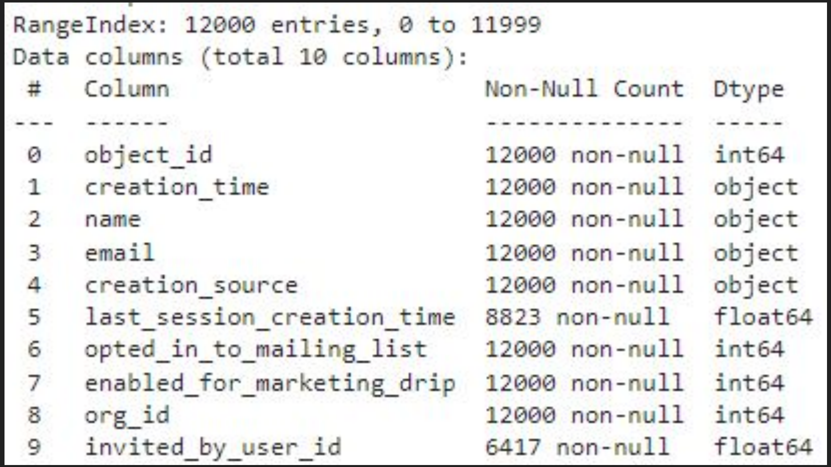
Predicting user behavior for customer service and marketing initiatives:

- ➔ Identify adopted users among all Relax users
- ➔ Find out the key feature driving user adoption

Data

- User information table

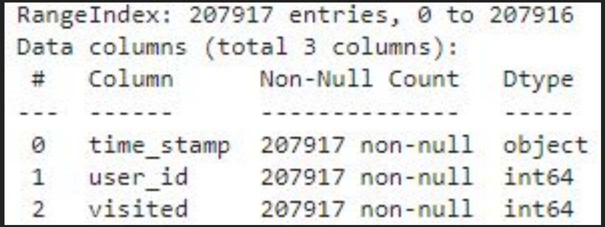
  (basic information)



```
RangeIndex: 12000 entries, 0 to 11999
Data columns (total 10 columns):
 #   Column                      Non-Null Count  Dtype  
---  ------                      --------------  -----  
 0   object_id                   12000 non-null  int64  
 1   creation_time               12000 non-null  object 
 2   name                        12000 non-null  object 
 3   email                       12000 non-null  object 
 4   creation_source             12000 non-null  object 
 5   last_session_creation_time  8823 non-null   float64
 6   opted_in_to_mailing_list    12000 non-null  int64  
 7   enabled_for_marketing_drip  12000 non-null  int64  
 8   org_id                      12000 non-null  int64  
 9   invited_by_user_id          6417 non-null   float64
```

- User engagement table

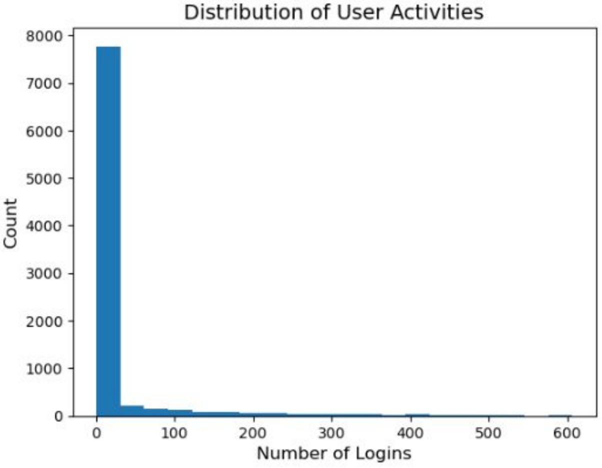
  (login time)



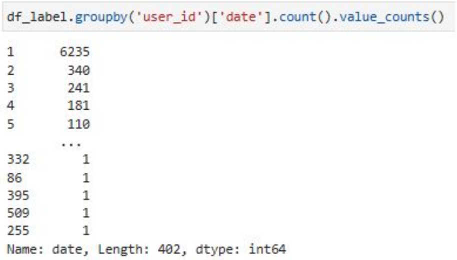
```
RangeIndex: 207917 entries, 0 to 207916
Data columns (total 3 columns):
 #   Column      Non-Null Count   Dtype 
---  ------      --------------   ----- 
 0   time_stamp  207917 non-null  object
 1   user_id     207917 non-null  int64 
 2   visited     207917 non-null  int64 
```

Adopted User Identification

- Loop over users with at least 3 logins
- Identify as adopted users if at least 3 logins within a 7-day period
  - 1,602 adopted users out of 12,000 users

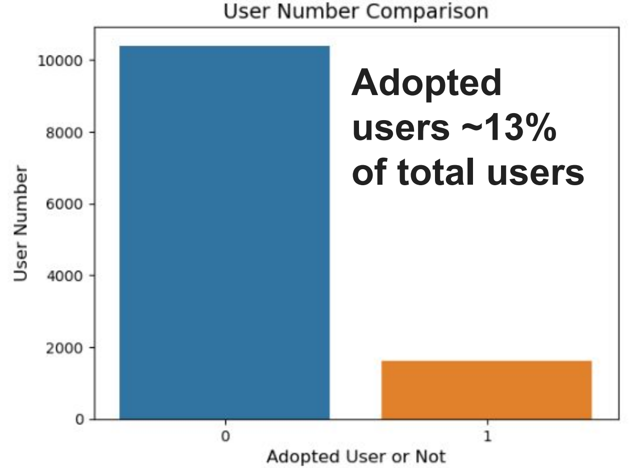


Half users only login once



```
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 users ~13% of total users

Data Wrangling

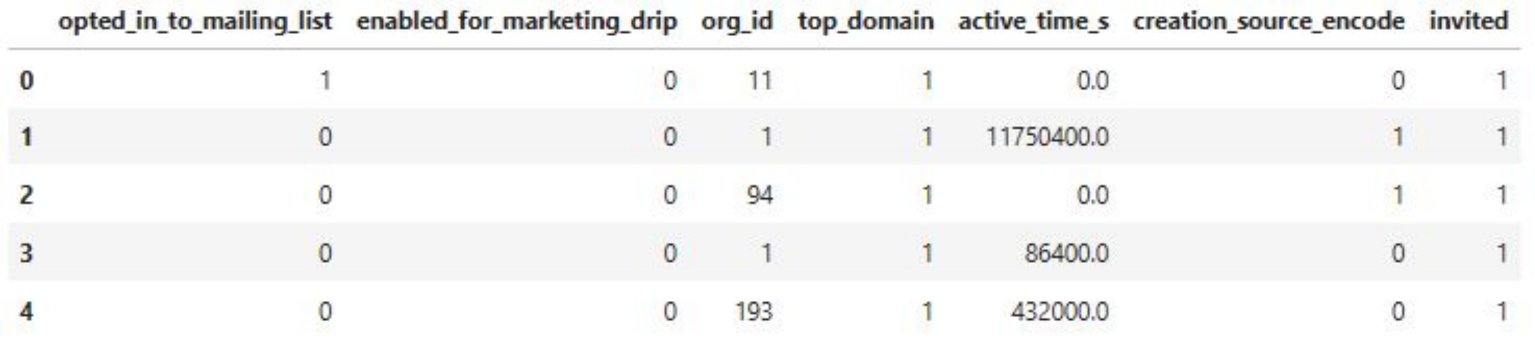
- Missing values:
  - `last_session_creation_time` has 3177 missing values
    - filled by `creation_time`
  - `invited_by_user_id` has 5583 missing values
    - filled by 0

- Categorical values:
  - Ordinary encoding for `creation_source`

Feature Engineering

- A binary indicator for **registered email** domain, top or not.
- A binary indicator for **invitation** by existing users.
- A numerical feature for **user active time** in seconds
  - 'last_session_creation_time' - 'creation_time'.

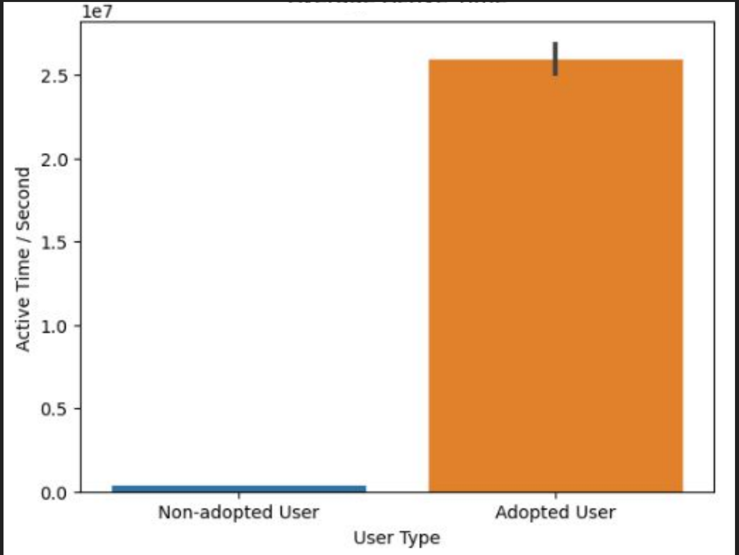
Feature Table

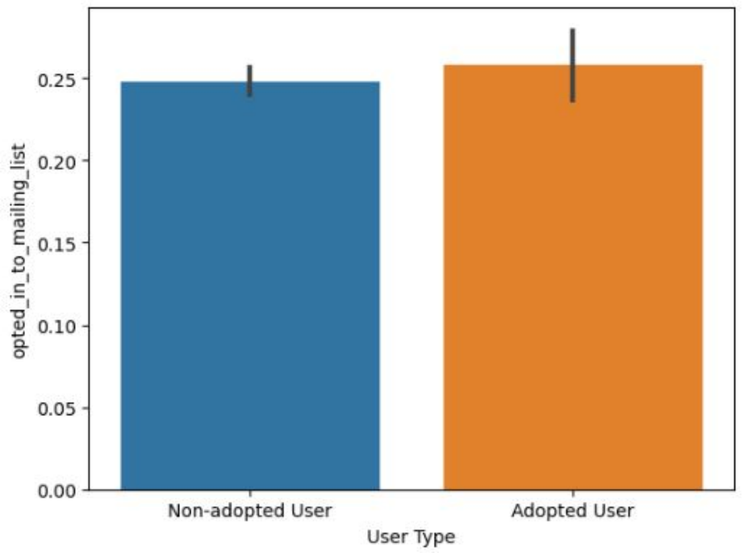


	opted_in_to_mailing_list	enabled_for_marketing_drip	org_id	top_domain	active_time_s	creation_source_encode	invited
0	1	0	11	1	0.0	0	1
1	0	0	1	1	11750400.0	1	1
2	0	0	94	1	0.0	1	1
3	0	0	1	1	86400.0	0	1
4	0	0	193	1	432000.0	0	1

EDA

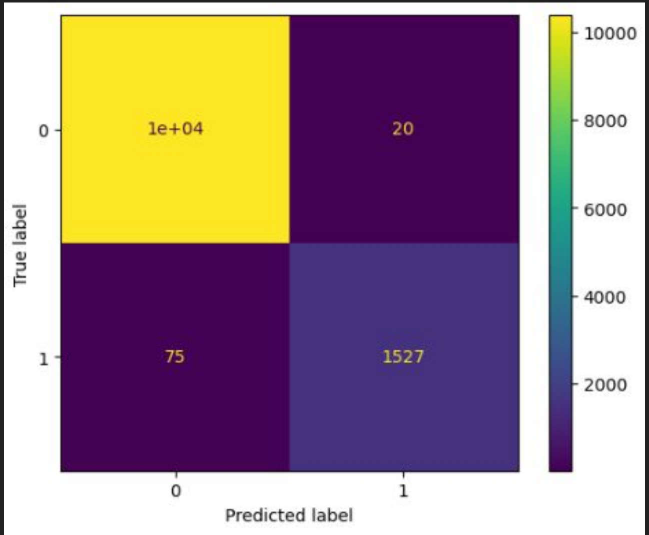
- The **average active time** is the only feature showing significant difference for adopted users and non-adopted users.



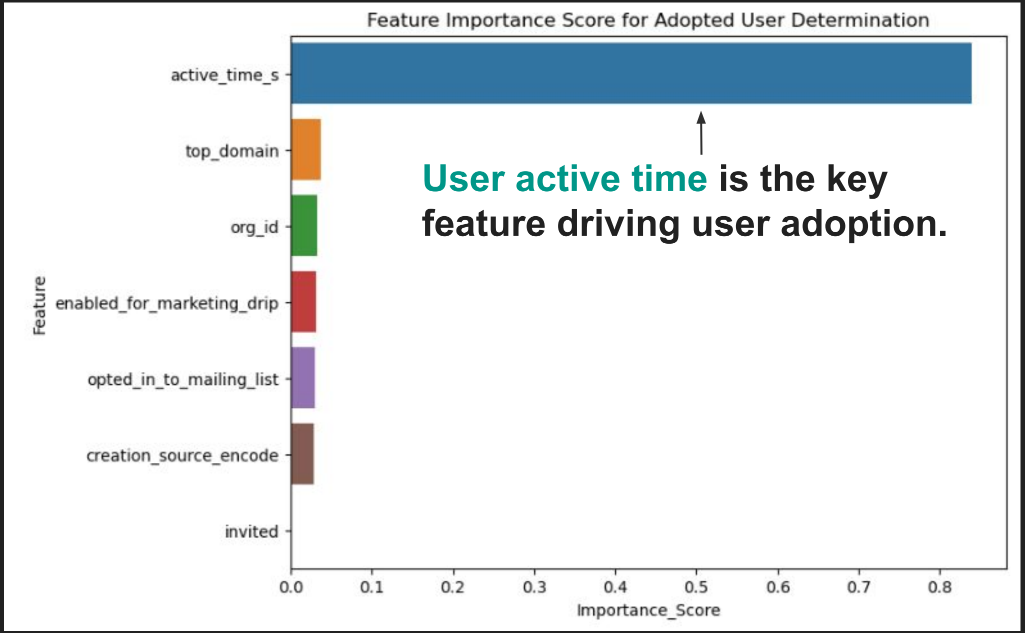


Modeling & Feature Importance

The XGBoost model is used for this binary classification task.



- **The model works good.**
- **The recall (false negative) needs to be further reduced.**



User active time is the key feature driving user adoption.

Summary

- Observations
  - Half of total users only login once after creating their accounts.
  - Adopted users only take ~13% of total users.
  - The most important feature determining if a user is adopted or not is the user's active time.
  - Adopted users tend to have much longer active time compared to non-adopted users.
- Business Suggestions
  - For upcoming marketing, Relax should focus on boosting user logins in order to increase user adoption.
- Future Work
  - Improve recall by hyper-parameter tuning and feature engineering
  - Create more user behavior related features, such as weekly/monthly login times