Relax Inc. Take-Home Challenge

Feature Engineering: I defined a a function as follows to create a new feature called **adopted**.

```python
def is_adopted(dates):
    '''Check if the user has 3 logins within any 7-day window'''
    sorted_dates = sorted(dates) # sorting the dates in ascending order.
    for n_dates in range(len(sorted_dates)-2): # need at least three logins.
        if (sorted_dates[n_dates+2]-sorted_dates[n_dates]).days <=7: # checks if three consequitive logins are within 7 days.
            return True
    return False
```

Feature Engineering and Machine Learning Model Development

Initially, the machine learning models constructed using the available features performed poorly. To improve model performance, additional features were engineered based on the existing data. The newly created features include **logins_within_first_month** (Number of logins within the first month of signup), **logins_within_first_week** (Number of logins within the first week of signup), **month_created** (The month in which the user signed up), **day_created** (The day of the week on which the user signed up), **days_since_signup** (The number of days since the user signed up).

After creating these features, appropriate imputation methods were applied to handle missing data. Subsequently, the dataset was divided into features and target variables, ensuring only relevant columns were included for modeling. For instance, columns such as the user's name and email, which are not expected to influence user behavior, were excluded. Conversely, features like the inviter, organization, and subscription to the mailing list, which can affect user behavior, were retained.

Model Development and Evaluation

Three machine learning models were developed using GridSearchCV to identify the best hyperparameters for each:

1. Logistic Regression : Accuracy: 0.9208, F1 Score: 0.7473, Precision: 0.6675, Recall: 0.8489

2. Gradient Boosting : Accuracy: 0.9258, F1 Score: 0.7608, Precision: 0.6852, Recall: 0.8550

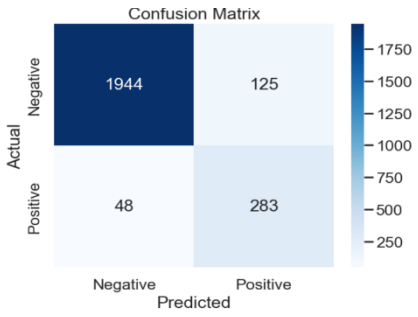
3. Random Forest : Accuracy: 0.9279, F1 Score: 0.7659, Precision: 0.6936, Recall: 0.8550

Key Insights

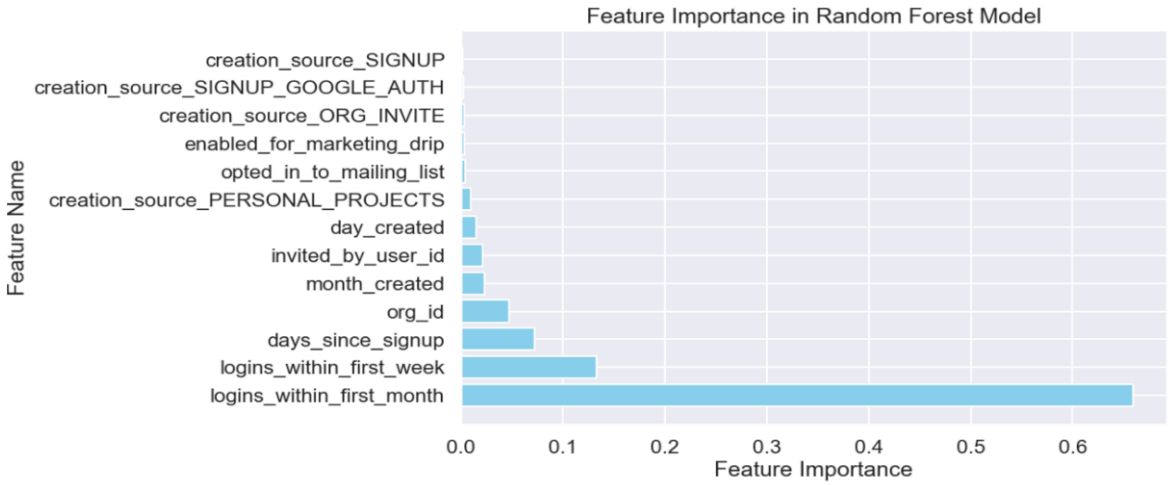
The **Random Forest** model outperformed the other models, achieving the best metrics across all evaluation criteria. Further analysis revealed that the features **logins_within_first_month** and **logins_within_first_week** were the top contributors in determining whether a user would be classified as "adopted."

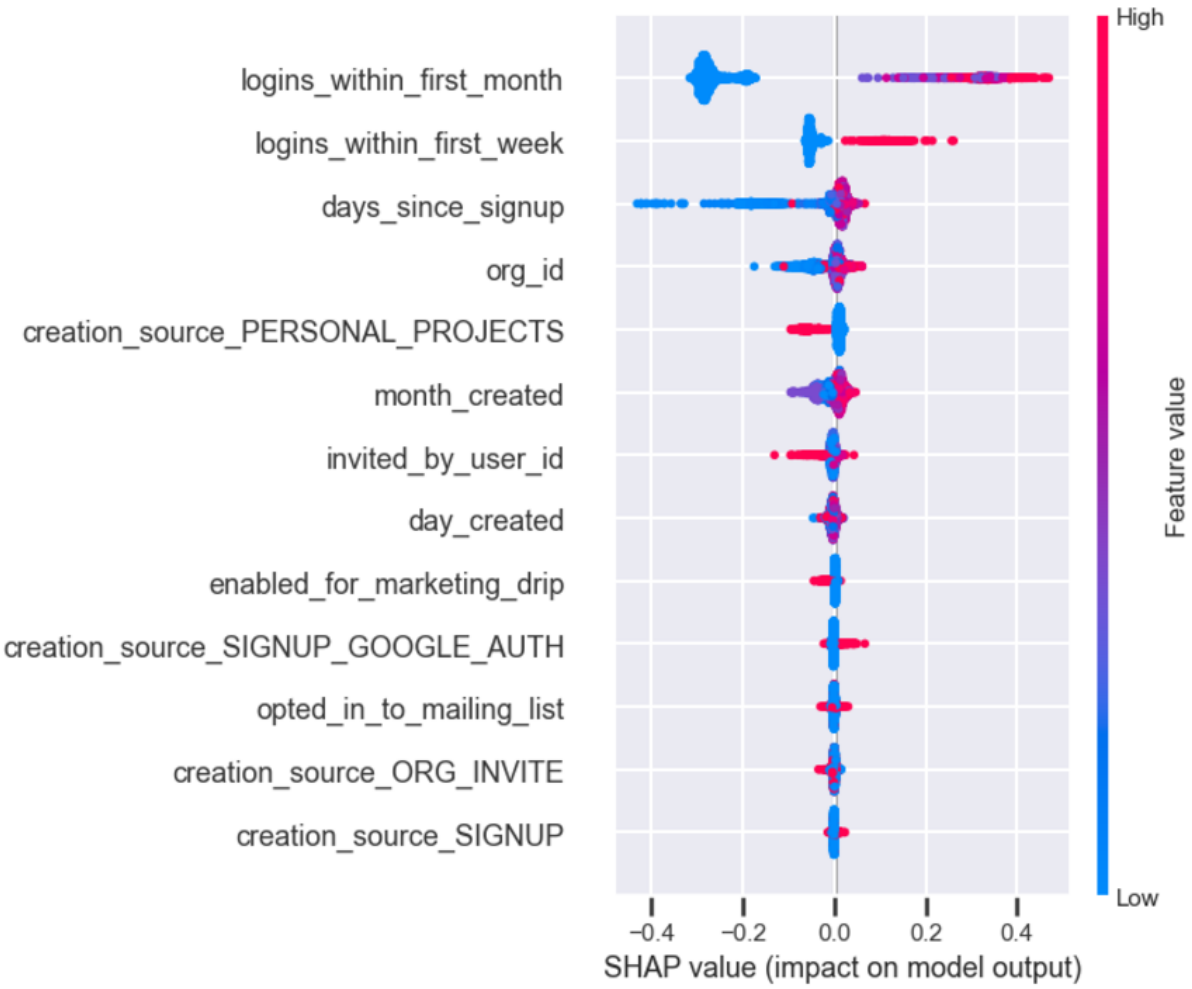
Recommendations

Given the importance of early user engagement, it is recommended that the company implement incentives to encourage users to remain active during their first month. This strategy can help establish a usage habit, increasing the likelihood of user retention and adoption.



Further details and analysis of this work are documented in the accompanying notebook within the same repository as this report.


Feature Importance in Random Forest Model
creation_source_SIGNUP
creation_source_SIGNUP_GOOGLE_AUTH
creation_source_ORG_INVITE
enabled_for_marketing_drip
opted_in_to_mailing_list
creation_source_PERSONAL_PROJECTS
day_created
invited_by_user_id
month_created
org_id
days_since_signup
logins_within_first_week
logins_within_first_month
Feature Name
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Feature Importance


logins_within_first_month
logins_within_first_week
days_since_signup
org_id
creation_source_PERSONAL_PROJECTS
month_created
invited_by_user_id
day_created
enabled_for_marketing_drip
creation_source_SIGNUP_GOOGLE_AUTH
opted_in_to_mailing_list
creation_source_ORG_INVITE
creation_source_SIGNUP
−0.4
−0.2
0.0
0.2
0.4
SHAP value (impact on model output)
High
Low
Feature value