The data provided from Relax Inc. was analyzed to determine which factors predict future user adoption. The code used to perform the analysis can be found in the jupyter notebook here.

The raw data was loaded from the two data files and prepared for analysis. The data was loaded, cleaned, merged, and transformed using Python. Feature and target variable arrays were created in preparation for the machine learning portion of the analysis. The target variable was created from the user engagement data. This binary target variable was assigned a value of 1 if the user was adopted, and 0 if the user was not adopted.

The following factors were used as features in the analysis:

- **creation_time_UNIX_ts:** UNIX timestamp of the user's account creation.
- **time_diff_from_creation_to_last_session**: difference in seconds between the user's account creation time and the user's last session creation time.
  NOTE -- Some users had missing values for the last session creation time, suggesting that they never logged in after creating an account. For each of these users, a sentinel value of -1 second was assigned to this feature.
- **org_id**: the organization the user belongs to, expressed as an integer organization ID.
- **enabled_for_marketing_drip**: binary variable indicating whether the user is on the regular marketing email drip. 1 - yes, 0 - no.
- **creation_source**: categorical variable indicating how the user's account was created. The creation source takes one of 5 values, expressed as a string.
- **opted_in_to_mailing_list**: binary variable indicating whether the user opted into receiving marketing emails. 1 - yes, 0 - no.
- **domain_category**: categorical variable indicating the domain of the user's email address. The domain category takes one of 7 values, expressed as a string.

See the jupyter notebook for more information about the motivation to use these features.

The feature importances were assessed by using a random forest classifier. The random forest was chosen over other models such as logistic regression because of its ability to capture the potentially complicated relationships between the variables with minimal feature engineering. The importance of each feature was evaluated using the mean decrease in accuracy with stratified k-fold cross validation. The mean decrease in accuracy was selected over the mean decrease in impurity because the heterogeneous data set contains categorical variables with varying cardinality, and the mean decrease in impurity could give biased results.

The results indicate that one factor in particular was practically significant for predicting user adoption. The time difference between the user's account creation and their last login had relatively high importance compared to the other factors. This factor alone should provide sufficient information to predict user adoption. The factor with the second largest importance was the account creation time. If desired, this factor could be included when predicting user adoption, although its importance was relatively low compared to the time difference factor.

Further analysis is recommended. Time-series analysis, including time-series segmentation, is recommended to evaluate the trends in user adoption over time in more detail.

Tables and Figures

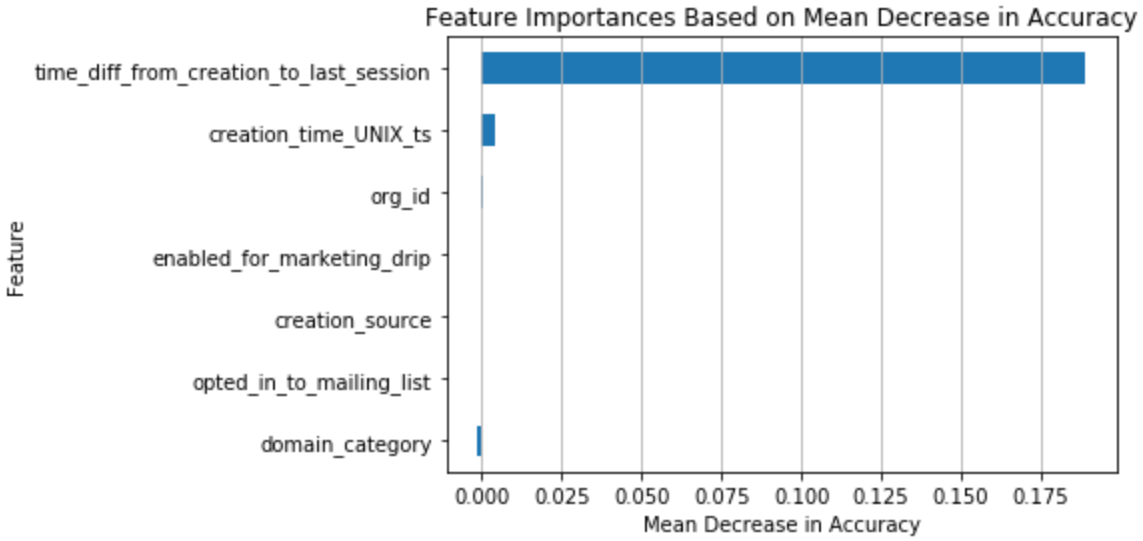

Figure 1: Feature importances for the factors used to predict user adoption.

Table 1: Mean decrease in accuracy for the factors used to predict user adoption.

	Mean Decrease in Accuracy
time_diff_from_creation_to_last_session	0.189000
creation_time_UNIX_ts	0.004000
org_id	0.000083
enabled_for_marketing_drip	-0.000084
creation_source	-0.000167
opted_in_to_mailing_list	-0.000250
domain_category	-0.001583