Relax Data Science Challenge Report

1. Introduction

This report details the analysis performed to identify factors predicting future user adoption for the product, as defined by the Relax Data Science Challenge. The primary goal is to understand what distinguishes an "adopted user" from others and to identify key features that predict this adoption. An adopted user is defined as a user who has logged into the product on three separate days in at least one seven-day period.

2. Data and Methodology

Two datasets were provided for this analysis:

- `takehome_users.csv`: Contains information about 12,000 users who signed up for the product, including their name, object ID, email, creation source, creation time, last session creation time, mailing list preferences, marketing drip status, organization ID, and invited by user ID.
- `takehome_user_engagement.csv`: Records daily login events for users, with a row for each day a user logged into the product.

Methodology

The analysis followed these steps:

1. **Data Loading and Preprocessing**: Both CSV files were loaded into pandas DataFrames. Timestamp columns were converted to datetime objects for easier manipulation. Missing values in `last_session_creation_time` and `invited_by_user_id` were handled appropriately.
2. **Defining Adopted Users**: A function was implemented to identify "adopted users" based on the provided definition: a user who has logged into the product on three separate days in at least one seven-day period. This involved grouping

user engagement data and checking login patterns within rolling seven-day windows.

3. **Feature Engineering**: Several new features were engineered from the existing data to enhance the predictive power of the model:
    - `account_age`: The duration in days between a user's creation time and their last session creation time. This aims to capture how long a user has been active on the platform.
    - `was_invited`: A binary indicator (1 or 0) showing whether a user was invited to join by another user.
    - One-hot encoding was applied to the `creation_source` categorical variable to convert it into a numerical format suitable for machine learning models.
4. **Model Training**: A RandomForestClassifier was chosen for its ability to handle both numerical and categorical features, its robustness to outliers, and its capacity to provide feature importances. The dataset was split into training and testing sets to evaluate the model's generalization performance. Due to the imbalanced nature of adopted vs. non-adopted users, `class_weight='balanced'` was used to prevent the model from being biased towards the majority class.
5. **Model Evaluation**: The model's performance was assessed using accuracy, precision, recall, and F1-score. A classification report was generated to provide a detailed breakdown of these metrics for both classes.
6. **Feature Importance Analysis**: The trained RandomForestClassifier provides a measure of feature importance, indicating the relative contribution of each feature to the model's predictions. This was crucial for identifying which factors are most predictive of user adoption.

3. Results and Findings

After performing the data analysis and training the RandomForestClassifier, the following key findings emerged regarding factors that predict future user adoption:

Model Performance

The model achieved an accuracy of approximately 95.8%. The classification report provides a more detailed view of the model's performance for each class:

```
              precision    recall  f1-score   support

         0.0       0.98      0.97      0.98      2080
         1.0       0.81      0.89      0.85       320

    accuracy                           0.96      2400
   macro avg       0.90      0.93      0.91      2400
weighted avg       0.96      0.96      0.96      2400
```

The model shows high precision and recall for predicting non-adopted users (class 0.0). For adopted users (class 1.0), the precision is 0.81 and recall is 0.89, indicating that the model is reasonably good at identifying adopted users, even with the class imbalance.

Feature Importances

The feature importance analysis revealed the following order of importance for predicting user adoption:

```
                                 importance
account_age                        0.981972
creation_source_PERSONAL_PROJECTS  0.004509
opted_in_to_mailing_list           0.004232
enabled_for_marketing_drip         0.002758
creation_source_GUEST_INVITE       0.001718
creation_source_SIGNUP_GOOGLE_AUTH 0.001561
creation_source_SIGNUP             0.001410
was_invited                        0.001022
creation_source_ORG_INVITE         0.000817
```

As evident from the feature importances, **`account_age`** is overwhelmingly the most significant predictor of user adoption, with an importance score of approximately 0.98. This suggests that the duration a user has been active on the platform (calculated as the difference between their last session and creation time) is a strong indicator of whether they will become an adopted user. Users who have been active for a longer period are much more likely to be adopted users.

Other features, while much less impactful than `account_age`, still contribute to the prediction. `creation_source_PERSONAL_PROJECTS`, `opted_in_to_mailing_list`, and

`enabled_for_marketing_drip` show some minor predictive power. This implies that users who were invited to personal projects, opted into mailing lists, or are on the marketing drip might have a slightly higher propensity for adoption, though their influence is minimal compared to `account_age`.

A visual representation of the feature importances can be found in `feature_importances.png`.

4. Conclusion and Recommendations

Conclusion: The most significant factor predicting user adoption is the `account_age`, which represents the duration a user has been active on the platform. This suggests that sustained engagement over time is a strong indicator of an adopted user. Other factors like `creation_source` and marketing preferences have a minor impact.

Recommendations:

1. **Focus on Early Engagement and Retention:** Since `account_age` is the primary predictor, efforts should be concentrated on encouraging new users to engage with the product consistently over time. This could involve:
    - **Improved Onboarding:** Streamline the initial user experience to ensure users quickly discover value and key features.
    - **Targeted Nudges and Reminders:** Implement smart notifications or emails to re-engage users who show signs of inactivity.
    - **Feature Highlighting:** Continuously showcase new or underutilized features to keep users engaged and exploring the product.
2. **Investigate `account_age` further:** While `account_age` is a strong predictor, understanding *why* users stay active for longer periods is crucial. Further research could involve:
    - **Qualitative Studies:** Conduct user interviews or surveys to understand the motivations and pain points of long-term users.
    - **Behavioral Analysis:** Analyze specific in-app behaviors of adopted users to identify patterns that lead to sustained engagement.
3. **Refine Marketing Strategies:** Although less impactful, the `creation_source` and marketing preferences still play a role. Tailoring marketing efforts based on

how users signed up (e.g., personal projects vs. direct signup) could yield marginal improvements in adoption rates.

4. **Consider Additional Data:** The current analysis is limited to the provided datasets. Incorporating additional data points could provide deeper insights, such as:
    - **In-app Usage Metrics:** Detailed data on feature usage, frequency of actions, and time spent within the application.
    - **Customer Support Interactions:** Understanding if support interactions correlate with adoption.
    - **User Feedback:** Direct feedback from users through surveys or feedback forms.

By focusing on fostering long-term engagement and continuously refining the user experience, the product can significantly increase its adopted user base.