Summary:

Aim of this case study is to check the variables which led to adopted users. At first glance, data is imbalanced with less adoptive users. Adoptive users were identified and various features were analyzed to see if they were impacting the user being adoptive or not.

Preprocessing Steps:

1. Data Understanding: Analyzed Data types, Missing values, Total user counts in both tables
2. Exploratory data analysis and cleaning: Changed the Data types of user id and date time columns.
3. From user_engagement data extracted the adopted users data by analyzing users who have logged into the application for 3 or more times in a 7 days period,for this we used .diff() and rolling method to identify the users fitting the given criteria.

Feature Engineering Steps:

1. Created a column named **adopted_user** in the user table created from user_engagement data.
2. Created a **user_referral** column indicating if a user was referred by someone and used data in the **invited_by_user_id** column.
3. Created dummy variables for **creation_source** column,
4. Did univariate and bivariate analysis and concluded that **org_id** column and **creation_source** column had relatively higher impact in a user being adoptive.

Model Selection:

Since adopted user data is imbalanced with most users being non adoptive , therefore, to balance the label variable which is adopted users, we used the imblearn and SMOTE technique in the train data to balance this variable.
Since this is a classification problem, we evaluated the Random Forest Classifier model as it gives us high accuracy, feature importance, robustness, reduces overfitting and is less sensitive to outliers.

Conclusion:

1. Most important feature was found to be org_id by both bivariate analysis and feature importance output by ML models.Followed by creation_source_PERSONAL_PROJECTS and creation_source_SIGNUP were found to be impacting adopted users a lot.
2. Model Results:
   Accuracy: Train -> ~83%, Test ->~75%
   ROC -> ~50%