

Summary :

- Once we create the column adopted_user, it is been observed that the user_ids being adopted user is less, hence class imbalance.
- There are missing/null values in the data.
- To handle class imbalance, we had to perform SMOTE.

Preprocessing Steps

- One hot encoding for the column - creation_source.
- Frequency/Distribution encoding for columns - org_id and invited_by_user_id since data in these columns is highly categorical.
- Changing columns to appropriate datatypes.
- Fill Missing Values appropriately.
- Handling Duplicates in data.

Model Selection

- Performed SMOTE for oversampling of minority class in training data set.
- Trained Random Forest model for checking the results.

Conclusion

- The adopted_user conversion is very less. In creation_source column we saw that ORG_INVITE followed by SIGNUP and GUEST_INVITE are important values so we can focus more here.
- Similarly we can use other column values link with adopted user (detailed discussion in notebook) for increasing adopted_users values.