Relax_data_science_challenge- Identifying Future User Adoption

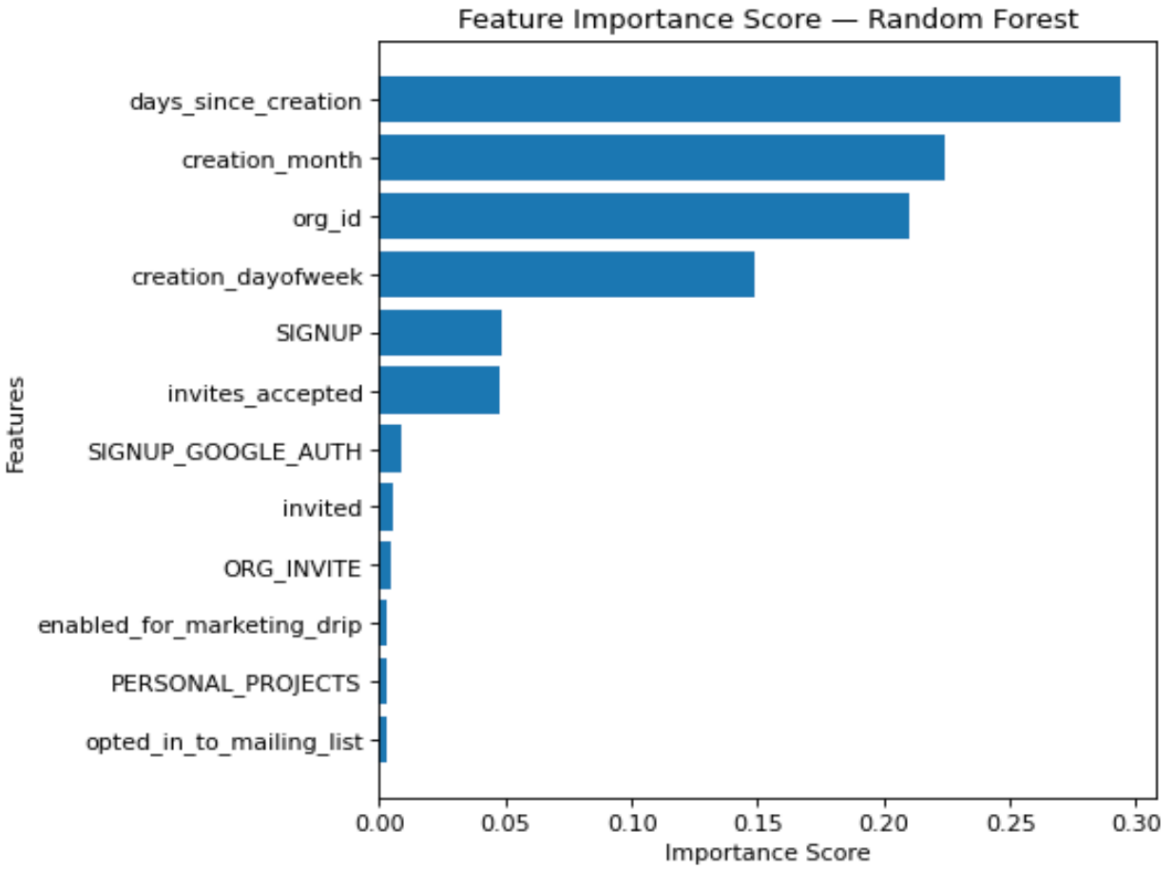
‘Adopted User” is defined as users who has logged into the product on three seperate days in at least one seven day period the first step was to generate the target feature from the given features.
With the engagement data set, user login times were aggregated to each day using groupby and to_datetime. Pd.rolling was used to get the rolling count of logins in a seven day window for each user login. The unique user_id with rolling counts above 3 were used to create the retention column in the users data set.

Feature engineering and cleaning -
1)The creation source was one hot encoded from the original five unique strings options.
2) To look for seasonality, creation time was used to create a retention by month.
3)As the invited by user id feature could not be used in its original form I generated an invitations accepted feature of the amount of times each user had someone else join due to their reference giving only 12 unique values but still encoding the information from the original invited by user column.
4) As the target was built around any 3 logins in 7 days and no the last seven days creation_time was converted to an integer of each date to be used as a feature.

Feature selection-
To take a look into the features correlation to retention a seaborn heatmap of the correlation between the columns was generated. Some of the one hot encoded features showed high correlation to the target but not each other which was a good sign for predictive power.

Training
A tree based model was selected to be able to show feature importance after training. Training a random forest model with max depth 5 and 100 estimators

The graph of feature importance shows days since creation and creation month and org_id as key factors



Days since creation , creation month and creation day of the week are all generated from the same initial feature pointing towards more features required to help build a more accurate model.