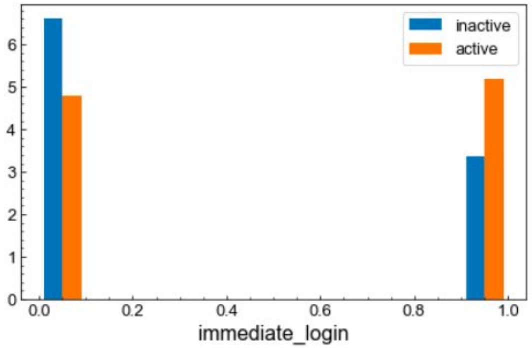
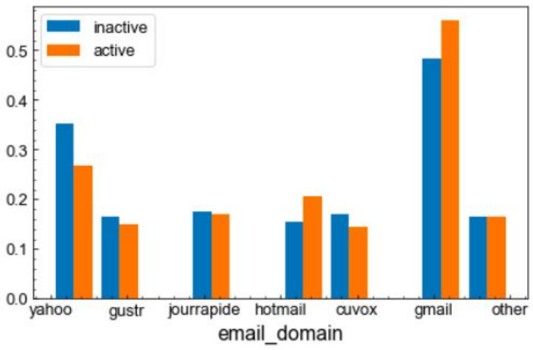
Features included in learning:

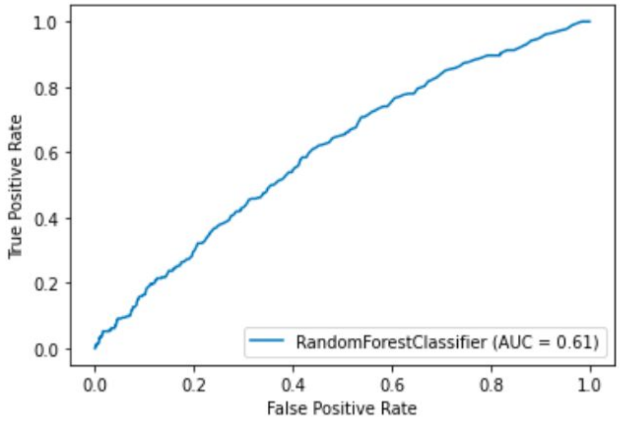
- Already provided features:
    - Whether user opted into mailing list (*opted_in_to_mailing_list*)
    - Whether user enabled their account for marketing drip (*enabled_for_marketing_drip*)
    - User's organization ID (*org_id*)
    - The ID of who invited the user, if anyone (*invited_by_user_id*)
    - The account creation source (*creation_source_SOURCE*)
- Modified or created features:
    - Whether the user has invited another user (*inviter*)
    - The month and year of account creation (*creation_month* and *creation_year*)
    - Email domain (*email_domain_DOMAIN*)
    - Whether the user's first login matches their account creation time (*immediate_login*)

From exploratory plots of the normalized feature distributions for active and non-active users there seems to be very little correlation between these features and a user's status. The email domain plot, provided as an example, is typical of the variance seen between the distributions. It is possible that there are some real effects from the *immediate_login* feature, or perhaps that users who create an account in April are more likely to become inactive.





Two types of tree-based methods, random forest and XGBoost, were optimized using balanced accuracy as a scoring metric. Both classifiers showed similarly poor performance but the random forest ensemble was slightly better with a balanced test accuracy of 51% and an ROC-AUC of 0.61 on the test set. The model feature importance was calculated using the SHAP package and confirms what was noted in exploration, namely that immediate login has the greatest impact on model prediction. However, the real take-away message here is that **these features do not have much predictive power for users' active status.** Further data collection and feature extraction would need to be performed to build such a model.


True Positive Rate
False Positive Rate
RandomForestClassifier (AUC = 0.61)