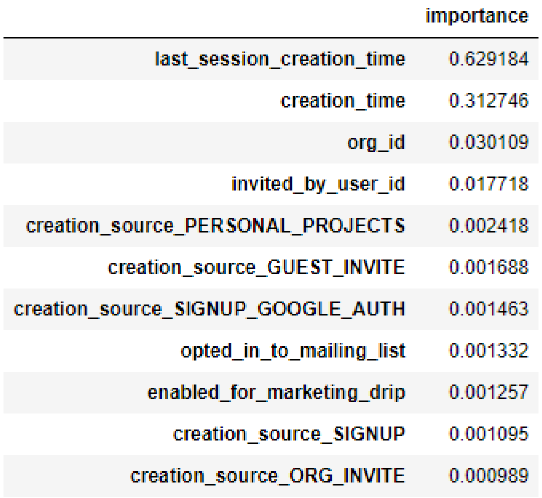
Gideon Blinick

Relax Challenge Write-up

Our task in this challenge was to “identify which factors predict future user adoption.” To that end, we fit a Random Forest to 70% of our dataset (training data), predicted on the remaining 30% (test data), and extracted feature importances from the model. We didn’t really need to split up our data since we were really only interested in discovering feature importance, but we did so because it is good practice. We also optimized for our model’s hyperparameters.

What we found was that we were able to achieve a very high accuracy of about 97% on the test set. When we looked at feature importances, here is what we found:



	importance
last_session_creation_time	0.629184
creation_time	0.312746
org_id	0.030109
invited_by_user_id	0.017718
creation_source_PERSONAL_PROJECTS	0.002418
creation_source_GUEST_INVITE	0.001688
creation_source_SIGNUP_GOOGLE_AUTH	0.001463
opted_in_to_mailing_list	0.001332
enabled_for_marketing_drip	0.001257
creation_source_SIGNUP	0.001095
creation_source_ORG_INVITE	0.000989

There it is. It’s not rocket science at all. Our most important feature is last_session_creation_time, or the variable indicating when a user last logged onto our product. It almost feels like cheating to use this variable in our model, even though this feature was not in any way involved in the creation of our “adopted” column. The reason this feature is so important is simple. The column contained many nulls which represented users that had never logged into the product. As part of our preprocessing, we set these nulls to 0 so we could feed it to scikit-learn. 0 was chosen because it was different from every other value and because it was so much smaller than every other value that the algorithm could interpret it as a login ‘very far in the past’. This feature was so important that when we removed it and re-ran the algorithm again, our accuracy declined to 81.3%.

We also see that the creation_time feature is important. This feature represents the time when a user created their account. When examining the correlation of it with our adopted target column, we found a negative correlation. A logical reason for this is that users that have been around for longer have had more time to become adopted users.

The other features are not nearly as important by comparison. Indeed almost 95% of the feature importance is due to the first 2 features in the table.

The code for the analysis, including the creating of our target column (‘adopted’) can be found in the same repository as this file.