**Task** Defining an "adopted user" as a user who has logged into the product on three separate days in at least one seven day period, identify which factors predict future user adoption .

**Data:** Two tables one with user engagement and the other table with user characteristics.

Step 1- Identify adopted users from user engagement table. Merge list of identified adopted users to user characteristics table. (See code for details)

Step 2- Clean user characteristics table.

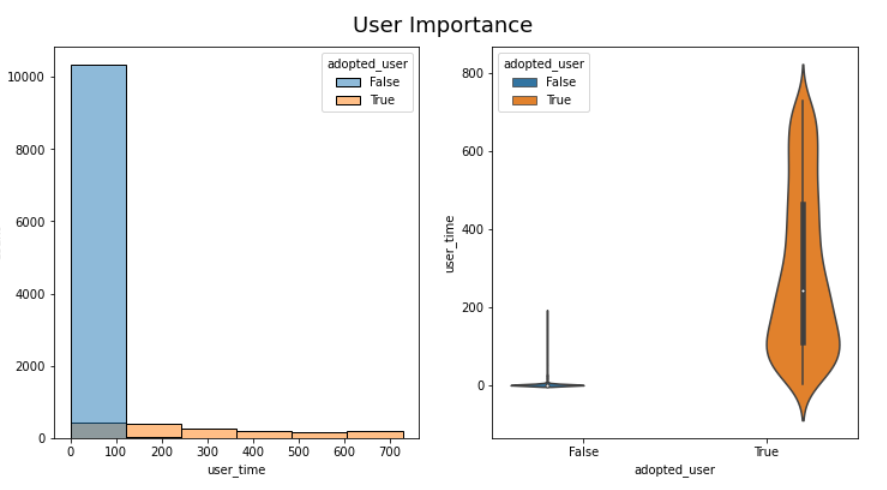
The table came with 9 columns and 12000 users. Two of the columns were time variables (creation\_time and last\_session). There were also null values in the last\_session column. I extracted the month and year from both time stamps, creating four new categories. This allowed us to have the data of which accounts were created when and if that had any value on adopted values. But by doing this we would lose how long the user had been around. So I also created a category called user\_time which was the number of days between the user’s last session time from creation time. This also allowed us to get rid of null values since if no last session time was listed I could say the user time was 0 since I would have to assume it was the same as start time. I then could drop both time variables.

Step 3- Feature Manipulation

Of the nine original features was email which had over 1100 unique variables. However emails come with three separate parts account name, subdomain and domain. For example for user [relax89@yahoo.com](mailto:relax89@yahoo.com) : account name (ex:relaxrocks89) , subdomain (yahoo) and domain (.com). Subdomains/domains can both divide the data and could tell us a lot about the user (.de domain for example suggests the user is in Germany). There were hundreds of sub domains so I limited it to domains that had over 10 users in for it.

Step 5- Machine Learning

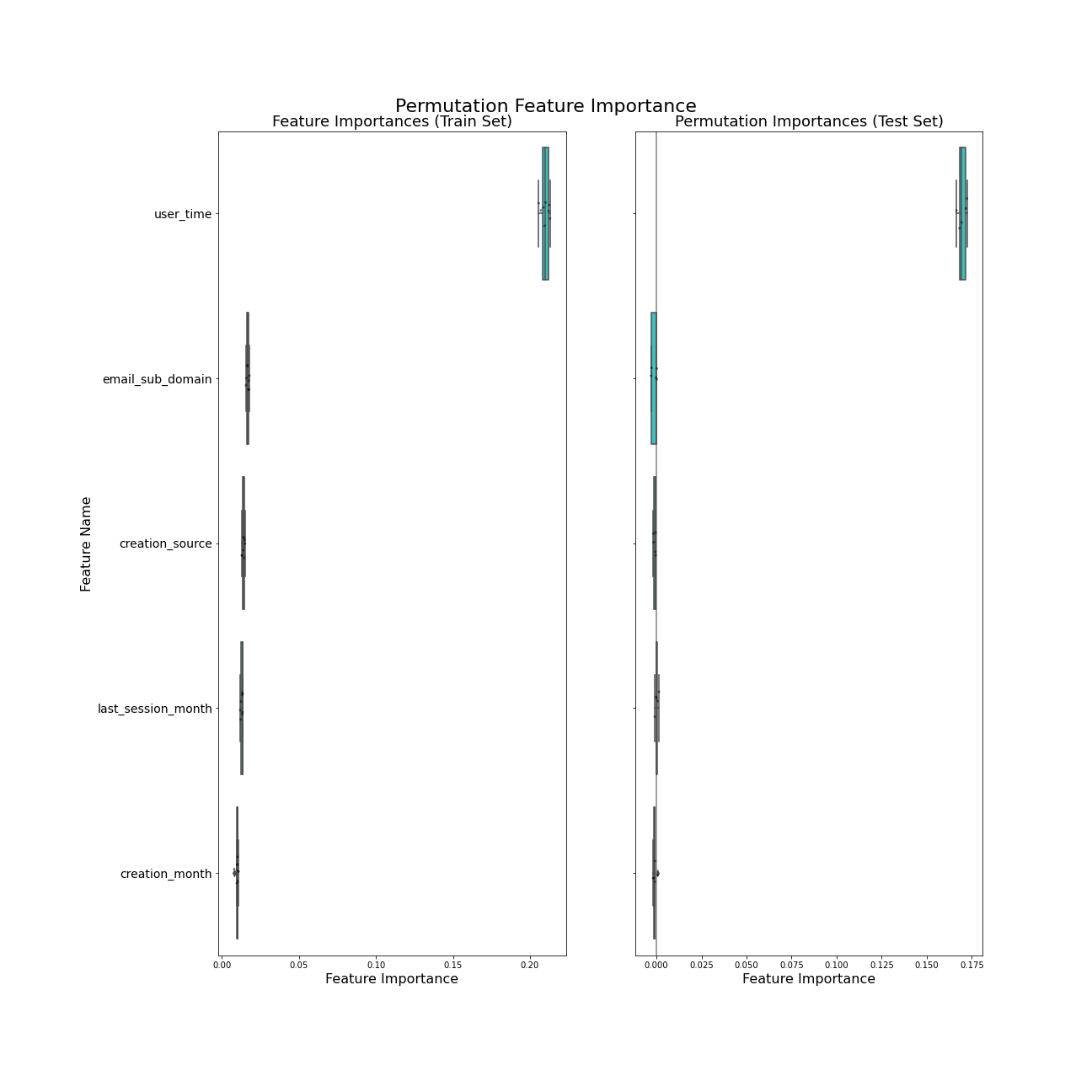
Having created these features I split my data into 60% train and 40% test, encoded my categorical features and ran it through a random forest classifier. My random forest came with 85% recall score and .97% accuracy score on the test data. When looking at feature importance user\_time turns out to have the highest value

Step 6- Visualize Feature Importance

There is one variable that is by far the most important- and that is how long they have been using the account. 4 out 5 adopted users have a use date of over 100 days. No adopted user has a zero day value (which makes sense) but over 50% of the non adopted users do. On the flipside a 200 day value has 50% of the adopted users while no non adopted users do.

Step 7- Other Traits

Less significant traits are email sub domain, creation\_source and when the user last played. However these are not as statistically significant.



saw as important.

