Alvaro Hu

Springboard

10/15/2018

Relax Challenge

Just for the record, I did the first half of this without Internet Access, so if some things seem a bit rudimentary, it is because of that. So, I started out by checking the data and seeing if there were any missing values. Sure enough, there were some in 'last_session_creation_time' and 'invited_by_user_id'. I filled the missing values in the former with the mean and eventually just dropped the latter, for I did not see how this could be an effective predictor. I then used the engagements table to figure out how many sign-ins each user had, and counted them, saving the result in a table called 'visited'. I then filtered out all of the users with less than 3 logins, for it would be impossible for them to have 3 sign-ins in a week if they don't even have that many in total.

The next part gets a little iffy, but I am confident it works. I made a list of tuples that contained each login along with its respective user, and then a second list with the same format, but 2 entries ahead. So effectively if it is the same user, I can tell how many days apart the two entries are, since they are sorted by day. If there were ever a time that there were two entries apart with the same user id that were separated by less than 7 days, then the user was added to a dictionary with a 1 as its value. I turned this dictionary into a DataFrame, merged it into the initial table, and filled the NaNs with 0's, and that is how I obtained my 'adopted' columns.

For the prediction, I first used a Logistic Regression and made dummy variables of the creation source. This gave me poor results, since there were so many 0's and only 1/10 of the data was 1's. This was predicting all 0's, and actually gave me a decent accuracy score, but terrible precision and accuracy. I then moved onto a Gradient Boosting Classifier in the sklearn.ensemble package and this provided me with much better results. My conclusions were that the time since the last login and the amount of logins a user has are the best indicators for whether or not they will become adopted users.