Mateo Martinez
Springboard DSCT
12/1/2020

Relax Inc. Report

This dataset contains approximately 200,000 rows of data, with 3 columns. A timestamp for a login time for a user_id for our company's website, Relac Inc. Our mission is to define active users of our website, which is defined by a user who has logged into the site more than three times in a calendar week. From our analysis, there are over eight thousand active users. Our company would like to find predictive features for active users. The table below outlines some features that could be used to build a predictive classifier model.

Week user_id	number_active_weeks	total_logins	active_user	first_login	last_login	first_active_week	first_login_week
1	0	1.0	False	2014-04-22 03:53:30	2014-04-22 03:53:30	NaN	2014-04-21
2	1	14.0	True	2013-11-15 03:45:04	2014-03-31 03:45:04	6.0	2013-11-11
3	0	1.0	False	2013-03-19 23:14:52	2013-03-19 23:14:52	NaN	2013-03-18
4	0	1.0	False	2013-05-22 08:09:28	2013-05-22 08:09:28	NaN	2013-05-20
5	0	1.0	False	2013-01-22 10:14:20	2013-01-22 10:14:20	NaN	2013-01-21

Using the 'first_login_week' column and the 'first_active_week' column, we could calculate a coefficient for the diminishing likelihood for a user to become active for each week that transpires. This could be used for targeted advertising for new customers to prompt them to continue using the website shortly after their first login. We could also analyze for correlations between the hour and day of week with active_users to see if there is a trend of weekend or evening users that could be used by the marketing department to conduct specific social media blasts at those times. Similarly, the login time during the 'first_active_week' could be analyzed for daily and hourly trends, to see if there is a correlation of day or time that users became frequent active users.

creation_source	last_session_creation_time	opted_in_to_mailing_list	enabled_for_marketing_drip	org_id	invited_by_user_id	active_user
GUEST_INVITE	1.398139e+09	1	0	11	10803.0	False
ORG_INVITE	1.396238e+09	0	0	1	316.0	True
ORG_INVITE	1.363735e+09	0	0	94	1525.0	False
GUEST_INVITE	1.369210e+09	0	0	1	5151.0	True
GUEST_INVITE	1.358850e+09	0	0	193	5240.0	False

Analyzing the signup metadata, this could be combined with our target feature of active_users to determine correlations and predictive coefficients between users that opted in for mailing lists, marketing drip and were either invited by a guest or by the organization. I would recommend using a logistic regression classifier using L1 and L2 regularization hyperparameters, and then analyzing the coefficients of that model. Additionally I would run a random forest model and analyze the feature importance of that model, taking into account that those coefficients have less interpretability than that of the logistic regression model.