CASE: Causal ML -Effectiveness of a Feature on an Online Dating Platform

Situation

More than half of single people have gone on a date with someone they met online and roughly a third of marriages in the US now begin online. However, online courtship norms are remarkably similar to patterns observed offline – men send out an abundance of messages with a low probability of finding a match, and women continue to be hesitant to make the first move.

Opportunity

One key feature of online dating is that the platform can help users interact. For example, the platform can act as a go-between and signal interest from users to each other. On the platform that we examine, a user can 'like' other profiles. The platform is wondering if it should allow users to see who has liked them. This could potentially increase engagement and the number of matches for users on the platform.

Key question

In order to assess this question, the platform designs a '*who likes you*' feature. As the name suggests, this allows users to see which other users have liked them. In order to assess if the feature increases engagement, the client designs an experiment (AB Test) where a treatment group gets to see other users who like them, and a control group of users does not have this ability. Many companies run AB tests. Standard statistics based hypothesis testing tells us whether group A (treatment) is, on average, different, from group B (control). We call this average treatment effect (ATE) and this can be dones using T-tests.

Of addtioanl interest here is to detect whether certain subroups (segments) of users (say women in the range of 37-42) have very high or very low treatment effects as a function of their X-vector characteristics. Technically we call this conditional average treatment effect (CATE) or heterogenous treatment effects (HTE) The big challenge is the we have to discover these segments, ie we don tno know in advance which sub-group is significnalty different in treatment versus control. This is where ML can help.

Key question

You need to assess if

(a) Users in the treatment and control groups are similar
(b) If the treatment (providing the Who likes you feature) increases the number of matches
(c) We will use the casual forest algorithm to estinmate the CATE and see where (for which ex ante unknown subgroup) the feature is particualty beneficial or hurtful.

Data

An online dating platform conducted an experiment on 11,172 new female users who signed up for free accounts on an online dating website. All these users joined in month 1 where none of them had the ability to see who liked them. In the second month, some of the users were given access to the *'who likes you'* feature.

Each observation in the data set is a single user. A match is defined as an exchange of at least four messages between a pair of users.

List of variables you will require for the analysis:

manipulation: Binary indicator if the user was in the treatment group and saw the WLY feature (value = 1) or in the control group (value = 0).

Dependent variable: You will need to create two columns in your table. The first column will be the sum of matches sent and received in month 1 (*match_rcvd_cnt_1 + match_sent_cnt_1*). Create a second column for the total matches in month 2.

Steps:

1) Import the data into R
2) Count the number of users in the treatment and control groups. Do we have roughly a 50% split?

Next, we are going to assess if the two groups are similar.

1) Calculate the averages for the age of the users in the two groups
2) Use the appropriate statistical test to see if the averages of the two groups are similar.
3) Can you see if the groups are similar across other dimensions? How about the total number of matches in the pre-treatment period (month 1)?

Next, we are going to assess if the effectiveness of the feature. There are two ways to do this:

1) Use the variable you have created for the total matches in month 2.
    a. What is the average matches for the control group?
    b. What is the difference between the treatment and control groups?

2) Using linear regression, regress the number of matches in month 2 on the indicator for if the user is in the treatment and control.
    a. What is the value of the intercept?
    b. What is the value of the coefficient?
    c. Is there a statistically significant increase in the number of matches for the treatment vs. control?
3) Finally, apply the causal forest algorithm to detect heterogenous treatment effects.