Workshop

Download the data from here:
https://docs.google.com/spreadsheets/d/1MfSSvX8KHqcYuQ__23HBY1DhsG9OgZAwgGiaAeOXwcE/edit?usp=sharing

The data is real product data however the features are disguised to obstruct company information. I've cleaned the dataset, normalized numeric variables and removed variables with near zero variance.

Product Data
retained - outcome variable, 1 denotes a purchase after 6 month, 0 does not
joined_january - feature, 1 denotes a january sign-up, 0 does not
joined_june - feature, 1 denoted a june sign-up, 0 does not
joined_december - feature, 1 denoted a december sign-up, 0 does not
category:Cards - feature, 1 denotes user sees a card prompt, 0 does not
category:Gifts - feature, 1 denotes user sees a gift prompt, 0 does not
category:Books - feature, 1 denotes user sees a book prompt, 0 does not
category:Paper_Products - feature, 1 denotes user sees a paper_product prompt, 0 does not
Size1 - feature, 1 denotes user seeing an ad image of size1 6X6, 0 does not
Size2 - feature, 1 denotes user seeing an ad image of size2 8X8, 0 does not
Size3 - feature, 1 denotes user seeing an ad image of size3 6X11, 0 does not
Size4 - feature, 1 denotes user seeing an ad image of size4 screen size, 0 does not
normalized_revenue - feature, numeric, user purchases total
normalized_quantity - feature, numeric, user number of purchases
Interaction - feature, numeric, interaction between revenue and quantity of purchases

1. Use any method to find which variables are the best predictors of the outcome variable Y. Which variable did you pick and why?

2. Let's try to find out if this factor is *causal*. I'll do a short tutorial of the '*Matching'* and '*rgenoud*' package.

a. Install the package '*Matching*.'

b. Try match balance between treated and control for your best predictor variable if it's a binary variable (if not, find a binary variable to test).
c. Use GenMatch to try to achieve balance…were you able to? (Hint: GenMatch grows exponentially in population size, 22k is a large dataset, subset your data, set a hard limit to the number of generations, probably around 5 or 10.)

d. Is this feature causal, why or why not?

e. Can you estimate the treatment effect of this variable.

Bonus: Individual Treatment Effects

Clearly this data set is observational and not randomly sampled without replacement. If we were able to achieve balance, we use causal inference trees to get individual treatment effects, instead of group treatment effects. Why is this important? We can predict who will benefit and how much they will benefit from a given treatment.

What is the effect of higher spend by users on individual treatment effects?

Use causal trees, an R package in development exists here: https://github.com/susanathey/causalTree and compares the effect to an uplift model from R package *uplift*. Note, that uplift models assume randomized treatment and control groups. What's the difference in the effect size on retention of higher spend that results between these two approaches?