Data Science and Strategic Pricing

Instructor: Jacob LaRiviere, Director of Economics and Data Science at Amazon

Emails: [jlarivi1@uw.edu](mailto:jlarivi1@uw.edu), [lghhager@uw.edu](mailto:lghhager@uw.edu) (TA)

# Course Assignments & Reading

Code, output and descriptive answers should be typed, printed and turned in at the start of class unless otherwise noted. Work in groups but everyone is required to turn in their own work with answers written in your own words. Keep in mind that in most cases a good answer is one precise sentence; quality is heavily favored over quantity. This will be graded on a full credit, half credit and no credit basis.

**Week 9, due Nov 29**

**Assignment to be turned in.** Please turn in your R output and answers to the questions.

1. Update your Double ML code to incorporate the regression tree classification you did to classify stores into three “leaves”.
   1. The code should start with classification (the regression tree) and then you should estimate a random forest/gradient boosted tree for stores in each leaf-brand-price and leaf-brand-quantity (18 forests, using our 3 leaf specification). You’ll then create a 3x3 elasticity matrix for each leaf (so three matrices).
   2. Make sure you have a robust set of features (e.g., make and use a bunch!) for use in building the random forests/gradient boosted tree.
   3. Compare the residuals from one of the forests/gradient boosted trees to a simple OLS model with the same RHS variables using a scatter plot. Does the Forest/gradient boosted tree beat OLS?
2. Suppose a firm offers a free trial for two months and has acquisition rate of 0.3, a conversion rate of 0.5, nobody buys the premium version directly (buy high rate=0), they make $10 per month if a customer converts, marginal costs per month are $5 and the average customer lifetime is 12 months. Suppose there are 1 million customers in their market.
   1. What is their total revenue currently?
   2. What are total costs (assume FC=0)?
   3. Based on this, what is gross margin (e.g., profits not accounting for fixed costs). Suppose fixed costs are 1 million. What is net margin (overall profits).
   4. They are considering reducing the length of the free trial to 1 month. Their data science team estimates the conversion rate will drop to 0.45 and the acquisition rate will drop to 0.28, what is the new revenue, new costs, and gross margins? Should they make the change?
3. Read the Online retail data into R. This is real anonymized purchase level data from a wholesaler in the UK.
   1. Calculate summary statistics including total number of unique customers, total number of unique countries, total revenue by country and average revenue by customer-country.
   2. In addition to averages plot the distribution of total revenue by customer using ggplot2 ([Plotting distributions (ggplot2) (cookbook-r.com)](http://www.cookbook-r.com/Graphs/Plotting_distributions_(ggplot2)/)).
   3. What percent of customers account for 80% of total revenue over the sample (Hint: calculate customer-level revenue, as above, sort from least to most revenue, use the cumsum() function in R, and use the output to compare to total revenue to get the percentiles).
   4. Define customer churn as having no purchases for three months. What is the monthly churn rate? Of customers who churn based upon that definition, what share of them purchase again?
   5. What is the LTV of a customer (in terms of revenue)?
   6. Can you identify returns in this data? If so, what percentage of orders are returned, and what fraction of total revenue is returned?
4. How does time of year impact order values?
   1. Summarize average number of sales per day and average revenue per day by month and year.
   2. Plot each on its own graph.
   3. The CEO of the company explains to you that they have the best ad strategy in the business. They only advertise close to the holidays (November and December) and those ads result in huge spikes in sales compared to when they’re not advertising. Do you agree with her? What would your response be?
5. We want to divide the customers into cohorts.
   1. Sort customers into “cohorts” based on the month and year of their earliest purchase (so all customers who first purchased today would be in a November 2023 cohort.
   2. How do metrics like number of orders, size of orders, number of returns, and size of returns vary across the cohorts?
   3. Using your evidence from (a) and (b), discuss the intuition of introducing subscription levels and what the impact on each cohort would be.