Newton, Sophia
ISYE6501x HW#10
Due July 25, 2018

Data set #3 (internal dataset)

Given {card type, product purchases, which pages, time spent per page, products clicked}
Use linear & logistic regression to determine if time spent, products clicked, card type are good predictors for item purchased or not.

Then the company could use this/these regression(s) to predictive model 1. Items to generate/create/produce/white label and sell themselves to edge out competitors (cough cough Amazon Basics cough cough) 2. Partner with sponsor companies to promote/sell their items 3. Then take the information of the items they sponsored/partnered with to compare sales of partner products, house products, unsponsored/non-house/unaffiliated products. Does sponsoring/featuring/making your own actually make money? Does burying unaffiliated products help users to select yours? Are users clicking through all 73 pages of “white lace dress”? Is this too close to home because I’m getting married in less than a month and I ordered my dress on the internet and had a very related to this homework set thing happen very specifically to me? Is this pretty closely related to the studies of filter-bubbling that people who don’t really understand the articles they read in Scientific American, were talking about a couple summers ago? All of these are probably yes, but a rigorous study that looks into it would say for sure.

Combine data sets 2 & 3 (credit bureau, internal set).

Given 2 {first name, middle name, last name, current city, real estate, monthly payment specific card, owed, paid, default status} and 3 {first name, middle initial, last name, card type, browsing history, time on page, clicks on page, purchase history}

First we want to clean this data, read: reconcile the 2 data sets. That’s a tough feat in and of itself but it’s one of my favorite parts of data analysis -- So we want to have a search (regex, grpl, contains, etc some kind of matching) to look for exact matches on all three, unaltered: 2FN:3FN, 2LN:3LN, 2cardtype:3cardtype. You also want to make your life easy, create a new variable in the second dataset: middle initial where you take just the first character of the middle name field. You also want it to have an exact match to the middle initial field of data set 3. So now we have a merged data set which has taken only those which have exact matches for all four of those fields, and it’s got all the information from those 2 datasets.

Now that we’ve reconciled the two datasets, what do we do with the mismatching data? The people who were in one but not the other, or who weren’t exact matches but were probable matches? Well honestly I’d probably to a probability match on the remaining, and if, for instance the levenshtein distance between two last name strings is particularly short (as it would be between Newton and Nweton - a likely typo) then you’re going to call that a probable match, and I’d put it into the merged data set but! I’d also add a new, categorical variable: match probability. You’d set that cutoff somewhere that you decided but I don’t have to set it right now. When your match is likely, you’re going to set it to 1. When your match is unlikely (levenshtein distance is LONG ie between Newton and Rudd) you’re going to set that probability to 0 - not a likely

match. And when all four are exact matches, you're pretttyyyyyy sure it's a match, you're going to set it to 2 very likely match. What about the situation where it's a Visa registered to John A. Smith? Well, that's a great question, I would probably need more information at that stage. However, if the name you're looking at is Stanley Albanowicz, well, I think regardless of the card type, it's going to be a probable match if all four match.

What are we doing with it once we've created this master index?

Marketing targeted products!! Am I going to market very long, wintery white, heavy fabric-ed dresses to the Sophia Newton who was looking for mid-length, light weight dresses if they live in Alaska? No probably not because I'm going to create new variables for median temperature and factor that into our suggestions too.

Is the person a real estate owner already? Then I'm probably not going to target my services as a consultant for helping first-time-home-buyers!

Are they already in default? Probably not going to market a product that costs $60,000 if they've got a $500 credit card in default.

That's what I'd do to monetize these data -- partner with other companies, do the research on maybe creating it myself, etc.