Homework 13

Due: Nov. 20, 2025

Question 19.1

Describe analytics models and data that could be used to make good recommendations to the retailer. How much shelf space should the company have, to maximize their sales or their profit?

Of course, there are some restrictions – for each product type, the retailer imposed a minimum amount of shelf space required, and a maximum amount that can be devoted; and of course, the physical size of each store means there's a total amount of shelf space that has to be used. But the key is the division of that shelf space among the product types.

For the purposes of this case, I want you to ignore other factors – for example, don't worry about promotions for certain products, and don't consider the fact that some companies pay stores to get more shelf space. Just think about the basic question asked by the retailer, and how you could use analytics to address it.

As part of your answer, I'd like you to think about how to *measure* the effects. How will you estimate the extra sales the company might get with different amounts of shelf space – and, for that matter, how will you determine whether the effect really exists at all? Maybe the retailer's hypotheses are not all true – can you use analytics to check?

What kind of data do you think you realistically have access to and can legally collect/use?

Introduction

This is a huge problem. Although I could not find solid numbers, Google's AI overview suggested that big box stores have 10K+ SKUs (stock keeping units), which aligns with grocery stores that typically have ~30K SKUs[1]. Therefore, we will want to mine as much preexisting data as we can.

There are three hypotheses we need to consider:

1. If a product is given more shelf space, then it will sell more units.
2. If there are more sales of a product, then there will be more sales of complementary products.
3. If complementary products are placed adjacent to one another, then the effects of #2 will be stronger.

Then using the relationships determined from 1 – 3, can we recommend an optimal shelving arrangement given practical constraints?

Notes

- Cross-validation – To avoid repetition, I will not continually state this, but given that this will be a huge dataset, and as we know, the more data thrown at a problem, the p-values can tend towards zero even if the correlation is spurious. It would be advisable to use k-fold cross-validation, or if the data is large enough, perhaps train on separate datasets and compare them.

Assumption of base data availability

I am going to assume for this study that the retailer keeps:

1. Transaction-level information on products sold. This matches what I have seen because some retailers can look up your transactions given your phone number or credit card.
2. The amount of shelf space each product is given. I am willing to make this assumption because the retailer is selling access, so they must be tracking that somehow.

[1] https://www.deloitte.com/us/en/insights/industry/retail-distribution/future-of-grocery-retail.html

Hypothesis 1: Shelf space's effect on sales

The relationship between shelf space and sales is obviously the basis for all the other analysis, so let's cover that first.

Given

- Transaction-level sales data,
- Product volume
- Packing density
- Product price relative to competitors,
- Brand type (store or name),
- Location within the store,
- Temporal factors (hour, day, week, and month all would have effects)

Use linear regression to estimate the sensitivity of sales to shelf space changes.

Hypothesis 2: Effect of complementary products

As stated in the Introduction, there are probably 10K+ SKUs for the retailer. This makes searching even for pairwise interactions a massive undertaking, with 50M two-pair combinations before even beginning to look at the factors that strengthen or weaken these interactions. Therefore, the first thing we need to do is narrow down to products with possible complementary sales. Given transaction-level sales records, use the definition of conditional probability: $p(A|B) = p(A \cap B)/p(B)$ to find the sensitivity relationship between two products as $p(n_i|n_j)$.

However, this full formulation is probably unnecessary in a retail setting where presumably most transactions are for small quantities and we can examine the simpler $p(i|j)$[2]. Then given the calculated probabilities $p(i)$ and $p(i|j)$, use the binomial tests to classify if to a statistically significant level

1. $p(i|j) > p(i)$ – we can call these complementary,
2. $p(i|j) \approx p(i)$ – we can call these non-complementary, or
3. $p(i|j) < p(i)$ – we can call these anti-complementary.

[2] The limitations of looking at pairwise interactions and only presence or absence might be problematic in some cases. For instance, in most normal circumstances Coke and Pepsi are anti-complementary; however, if accompanied by purchases of large amounts of chips, meat, side dishes, etc. they might be complementary because the customer is preparing for a large barbeque with people that prefer each.

Let's define the degree of complementariness as $c(i, j) = p(i|j) - p(i)$.

Hypothesis 3: Effect of proximity on degree of complementariness

From #2 we defined a metric called the degree of complementariness which measure how much more likely a product is going to be purchased given another product is purchased. The retailer obviously wants to increase sales and so wants to find ways to increase that quantity. If the retailer has information about where all the products are in the store we could simply rely on that data (and I know some retailers for example the Home Depot do list on their website where all products are). However, assuming the retailer does not have this information, we might need to create this data.

Given:

- Sets of products classified as complementary, non-complementary, and anti-complementary,
- Several sets of shelves varying from adjacent to far apart

Collect data and create a linear model to estimate the sales sensitivity to complementarity and distance. Since this is an experiment involving humans, a high amount of replication is required to average out confounders.

Store planning

Given information such as:

- Permissible store layouts (these would need to follow fire codes and other applicable laws, best practices regarding customer preferences,
- Product characteristics (profit margin, space requirements, packing density)
- Degree of complementariness between products
- Sales sensitivity to complementarity and distance
- Shelf space constraints (min/max bounds, total space available)

Use an optimization model (e.g., linear programming) to determine product locations and shelf space allocations that maximize store profit.