Rotman

Master of Management Analytics

Optimizing MM&A's In-Store Experience: Introducing the Dedicated Instabasket Aisle

Prepared by

Team E:

UofT Statistic: Yunrong Liu: yunrong.liu@mail.utoronto.ca

UofT Statistic & Economics: Dewei Lin: dewei.lin@mail.utoronto.ca

UofT Management, BBA: Yijun Gu: yijun.gu@mail.utoronto.ca

UofT Computer Science: Zheyuan Wei: zheyuan.wei2003@gmail.com

UofT Statistic & Economics: Huiyan Zhang: fionaaa.zhang@mail.utoronto.ca

UofT Statistic & Mathematics: Haoyang Han: jimmyhan1003@gmail.com

Find us on Linkedin

Rotman School of Management
UNIVERSITY OF TORONTO

Preview Of The Structure

Pre-analysis & Preparation of the data

Start From Business...
Page 03

What challenges are we currently facing? What solutions do we propose? What results do we anticipate?

Data Processing: Storage Condition & Finding Substitute
Page 04、05

The 'Refrigerated/Frozen' category is not classified correctly. How can we rectify this?
How can we find substitutes for our products?

Exploratory data analysis
Page 06 & 07

After cleaning and substitution, what does our dataset look like? Can we derive any insights or intuitions from it?

Model analysis & Result of the data

Modeling
Page 08~11

How do we model to obtain results?
How do we design our model?

Result Analysis
Page 12

What are our results and how should we interpret them?

Business Insight
Page 13

Getting back to business, what are our conclusions and future directions?

Start From Business...

Current situation

Supermarket

1. Changing business dynamics due to on-demand grocery delivery.
2. Separate checkout areas/processes for online orders.

Customer

Online:
1. Ability to shop from anywhere at any time.
2. Access to product reviews, ratings, and detailed descriptions.

In person:
1. Traditional shopping experience with familiarity to store layout and products.
2. Dependence on in-store promotions, displays, and staff interactions for information.

Existing problems

Supermarket

1. Product substitution and communication issues.
2. Potential aisle congestion during peak times.

Customer

Online:
1. Potential product substitutions.
2. Communication gaps in product availability.

In person:
Possible disruptions due to increased personal shoppers.

Solutions

1. Created a dedicated "Instabasket" aisle for online orders.
2. Aisle contains a max of 1,000 products; Upto 100 refrigerated and 100 frozen.
3. Identified substitutable products for unavailable items.
4. 5% discount offered for accepted product substitutes.

Expected results

Supermarket

1. Faster Order Processing: With most items readily available in one aisle, order preparation and delivery should be quicker.
2. Improved customer satisfaction

Customer

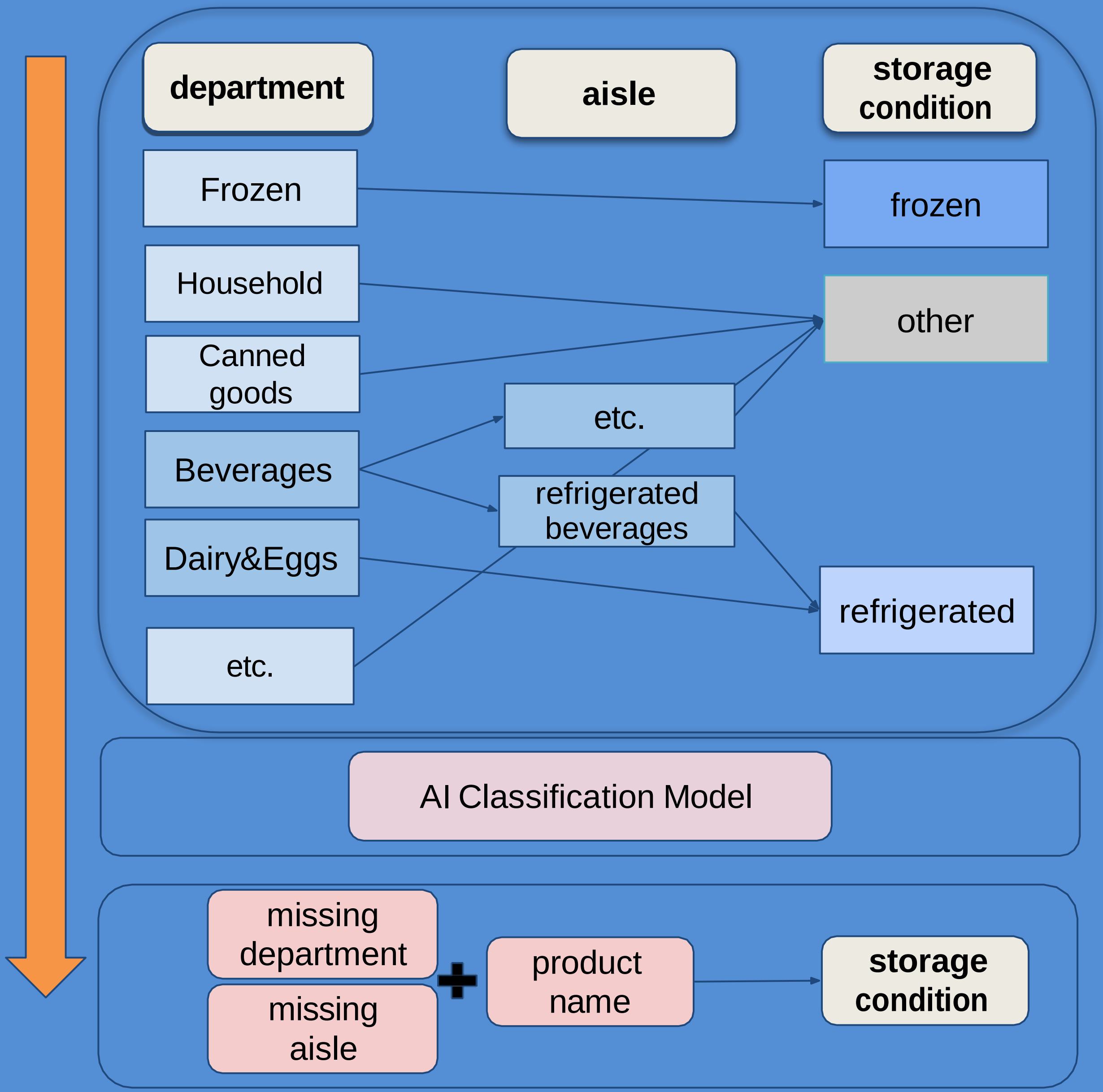
Online:
1. Obtain desired items more promptly.
2. Increased acceptance of alternative products.

In person:
Reduced store disruption:
With personal shoppers mainly operating within one aisle,in-person customers might find the store less crowded and more navigable

Rotman

Data Processing: Storage Conditions

Data Workflow



The classification of frozen and refrigerated products is inaccurate. We need to take action to rectify this.

Step 1: Classify up to 30% of the products based on department or aisle. As depicted in the plot, we've successfully categorized 30% of the total products into 'frozen' and 'refrigerated' sections. There are NO refrigerated department, we can only put dairy and egg in refrigerated in this stage.

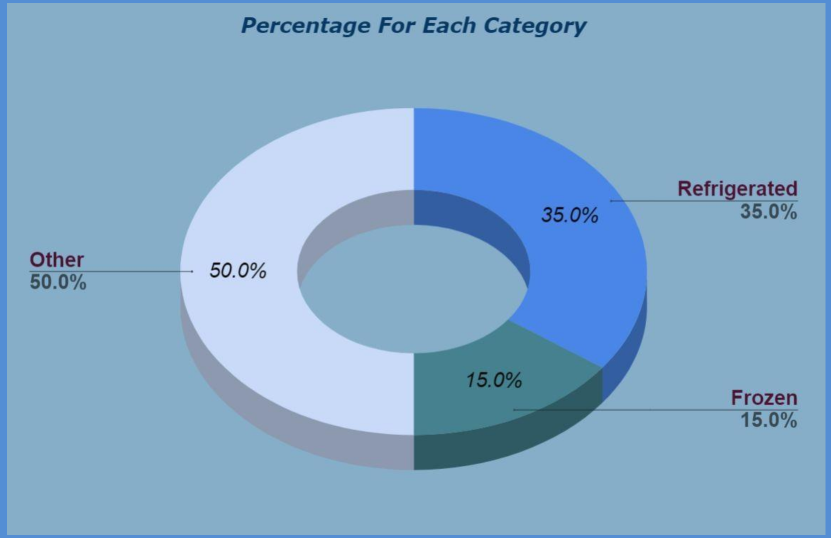
Step 2: Challenges: Extract the aforementioned 30% of products for classification. About 67% of products pose ambiguity in storage conditions, necessitating extensive human judgment. Some products either have multiple storage conditions or just one unclear condition. Over 3,000 products have non-informative aisle and department labels, such as "missing" or "others."

Step 3:We've developed an AI model trained on data classified as 'frozen' or 'refrigerated'. This model determines how products should be stored.

Step 4: Implement our AI model from Step 3 to classify the remaining 67% of the data.

Step 5: Post-classification, every product is labeled as either 'others,' 'refrigerated,' or 'frozen'. The result is shown below.

Step 6: To validate our classification: We randomly selected samples: 200 from 'others', 140 from 'refrigerated', and 60 from 'frozen'. Human evaluators checked these samples. Results confirmed that all sampled products were classified accurately.



Rotman

Finding Substitute: Before proceeding with modeling, we aim to address the issue of substitutes, they are essential for our metric calculations.

If this item becomes unavailable...

"Image sourced from Walmart's official website."

> Eg: Substitute Group:Milk

- Original Pure Creamy Almond Milk
- Organic Unsweetened Almond Milk
- Organic Vanilla Almond Milk
- Unsweetened Almondmilk
- Original Almondmilk

Toasted Coconut Almondmilk Blend

Organic Unsweetened Almond Milk	Refrigerated
Original Almondmilk	Refrigerated

Why they will pick substitutes?

Customers are satisfied with the **5% discount OR the product could not be found**

Why we want them to pick substitute?

Potential **more usage** for the 1000 products

How to find substitutes?

Using a technique called semantic approximation.

We developed a model that identifies words with similar meanings within the same department and aisle.

Category number :27

Result

Total Category: **4k+**

Average item in category: **8**

Total 1M

Products

300K

Trained phrases

35000 Unique Products

-> 4000 Categories

Now we have finish all of our data pre processing, it's time for:

Exploratory data analysis

Original Data: 987259 records, 100000 orders

Remove rows with text values in 'aisle_id' and 'department_id'

Remove the Empty column or erroneous data

Cleaned Data: 786406 records, 77936 orders

1. Total number of unique products: 33107

E.g. Hoisin Garlic Marinade & Sauce, Super Plus Security Tampons, Honey Raw Texas Wildflower

2. Total number of different aisles: 134

E.g. frozen meat seafood, packaged produce, fresh vegetables

3. Total number of different departments: 21

E.g. dairy eggs, produce, canned goods

4. Total number of 'Refrigerated', 'Frozen' products:

'Refrigerated' products account for 6,050 entries or 18.27%.

'Frozen' products constitute 2,988 entries, representing 9.03% of the total.

"'Other' category encompasses the majority with 24,069 entries, amounting to 72.70% of the overall dataset."

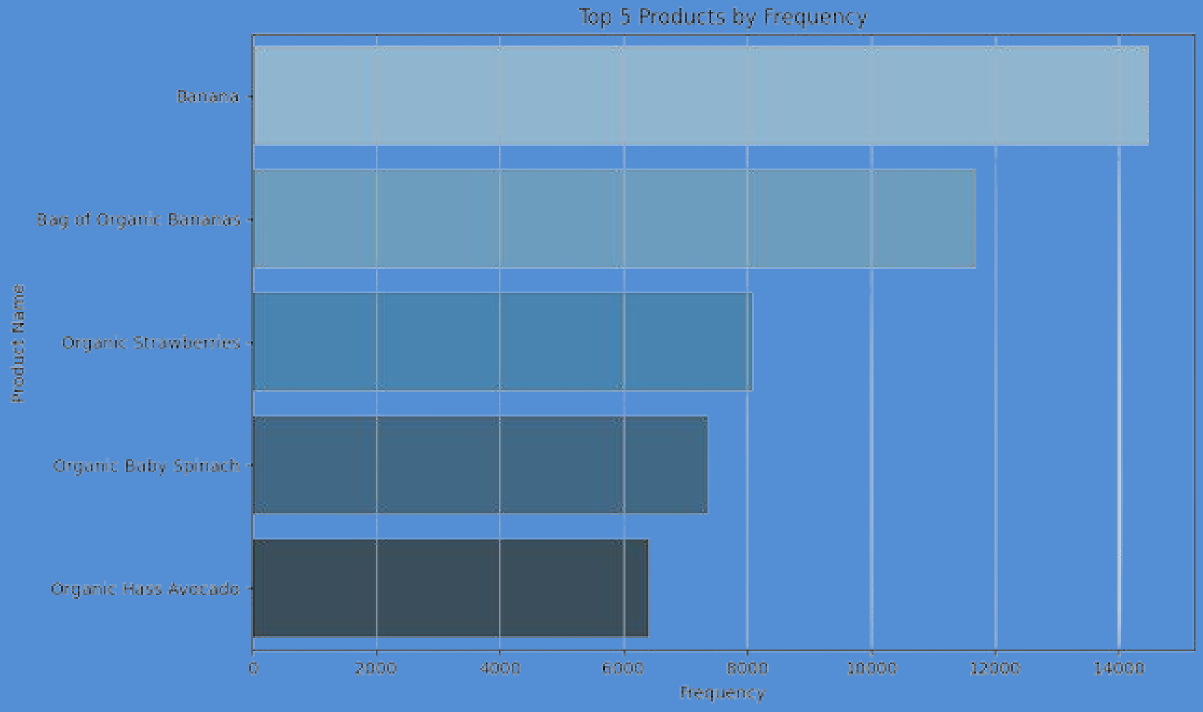
5. Total number of substitute groups: 4160

Average number of unique products per substitute group: 7.96

E.g. Data with a substitute_group value of 1000, where product_id and product_name are distinct (or different):
Chocolate Covered Biscuit Sticks, Original Fudge Sticks

Continue on exploratory data analysis...

The top items with the highest occurrences in all orders:



The product "Banana" with ID 24852 tops the list with 14,494 orders. "Bag of Organic Bananas" (ID 13176) follows with 11,694 orders. "Organic Strawberries" (ID 21137) has 8,081 orders, "Organic Baby Spinach" (ID 21903) has 7,369, and "Organic Hass Avocado" (ID 47209) has been ordered 6,411 times.

Frequency Analysis of Product Occurrence in 77,936 Orders:

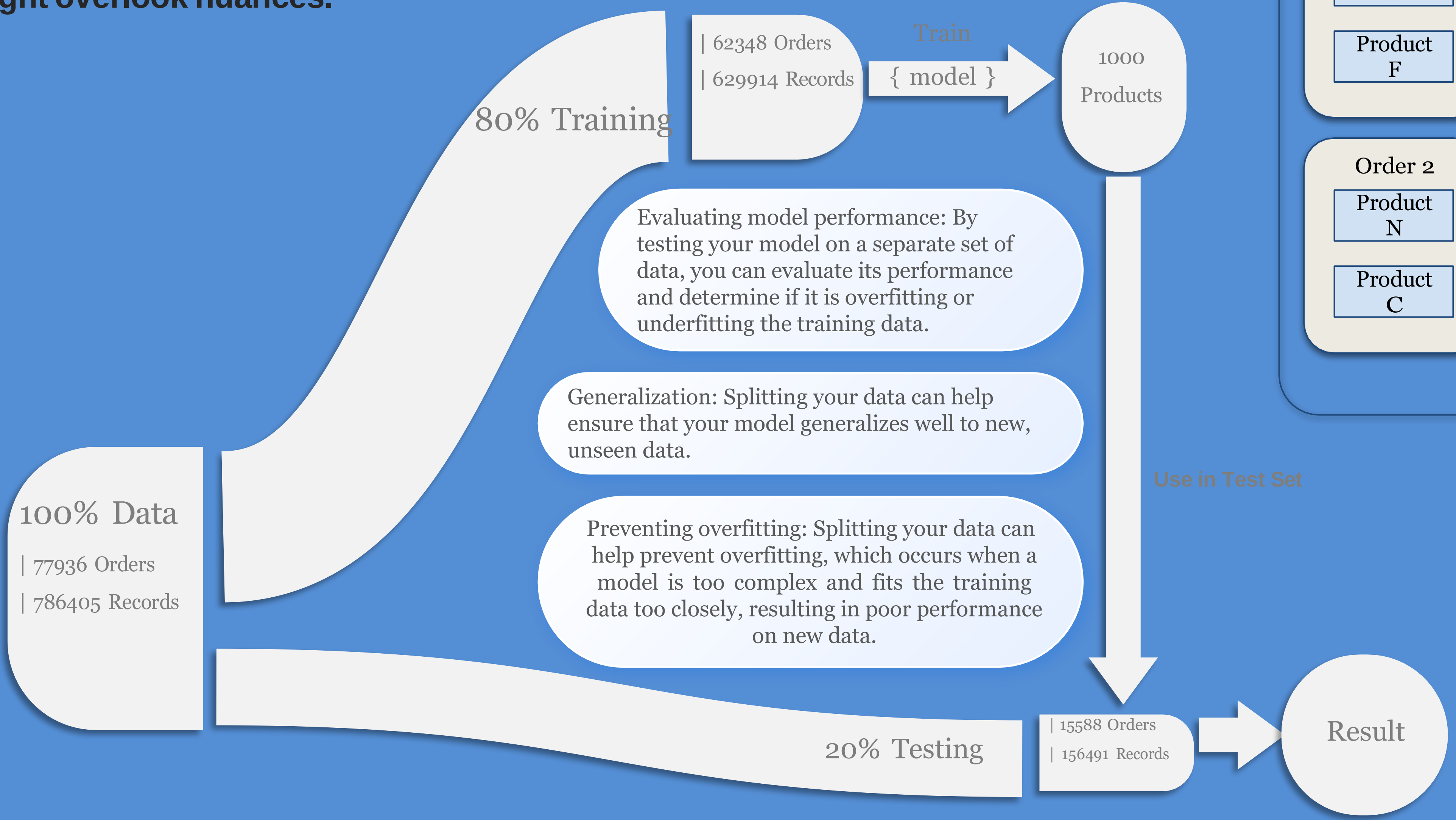
Product Occurrence	Number Of Products
Above 500	179
250 to 500	284
Below 250	32643

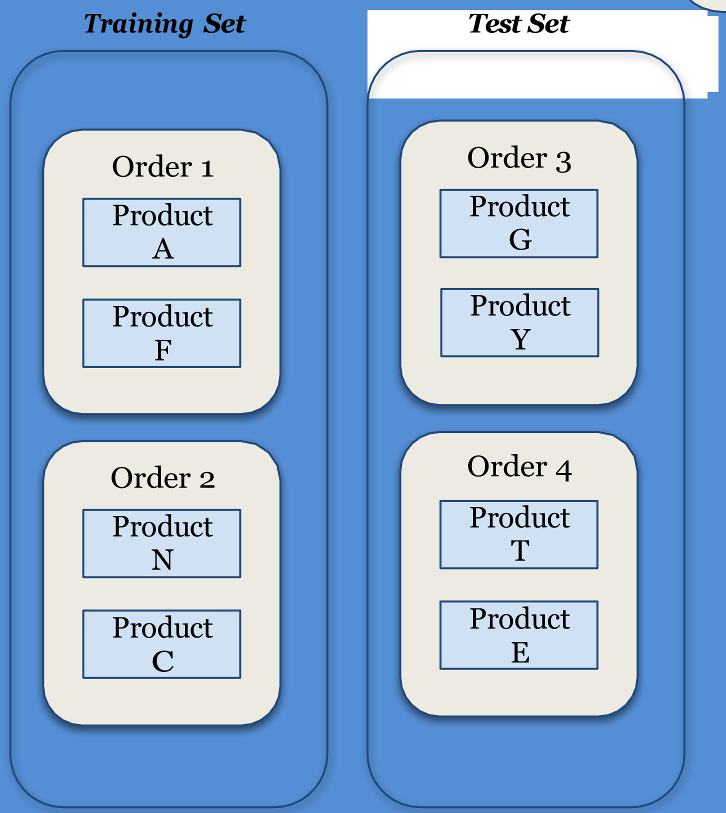
It can be seen that we have 179 products with a frequency of 500 times or more, and 284 products with a frequency between 250-500 times.

This indicates that the product reuse rate is very high. Therefore, selecting 1,000 products for a Instabasket Aisle for online orders can cover a significant portion of order requirements.

Preparing to Modeling

We consider every single order is a group, based on that, we split the data. This method ensures consistency and granularity in our analysis, allowing us to obtain detailed insights specific to each order rather than generalized patterns that might overlook nuances.





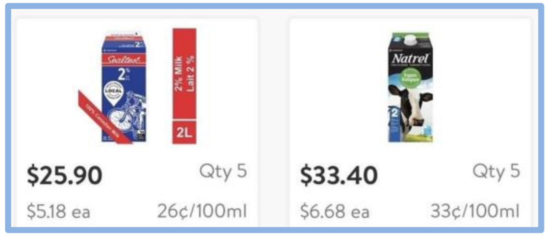
Apply Substitution:

The analysis below outlines the rationale behind our decision to apply substitutes solely in the context of a 5% discount.

Customers may require substitutes under two primary circumstances:
1: When they accept the 5% discount offer.
2: When the desired product is unavailable.
The latter scenario doesn't impact our final results. We cannot dictate customer purchase choices; rather, our role is to suggest alternatives by offering incentives, such as discounts.
If a customer chooses a product from the 'Instabasket' and it's unavailable, they might opt for a product outside of this list. However, we lack data on such occurrences, so it would be speculative to assume a frequency for this scenario. Thus, we've decided not to factor it into our analysis.
On the other hand, if a customer selects an item not listed in the 'Instabasket', and even if there's a substitute available within the 'Instabasket', they might still look outside of the aisle. This is because, despite the substitutability of products, they remain distinct entities. Thus, an internal substitute within the "Instabasket" won't influence an external purchase unless the 5% discount is in play."

METHOD

- Every time choose products not in Instabasket, we recommend products in the Instabasket, with 5% discount



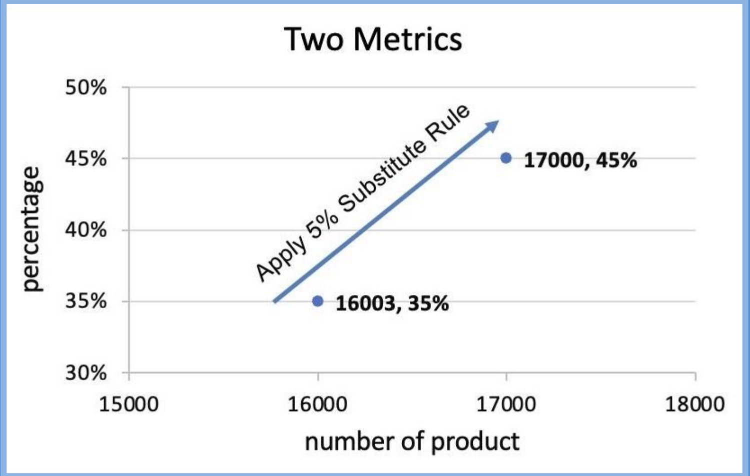
Not in Instabasket

In Instabasket

RESULT

We assume:

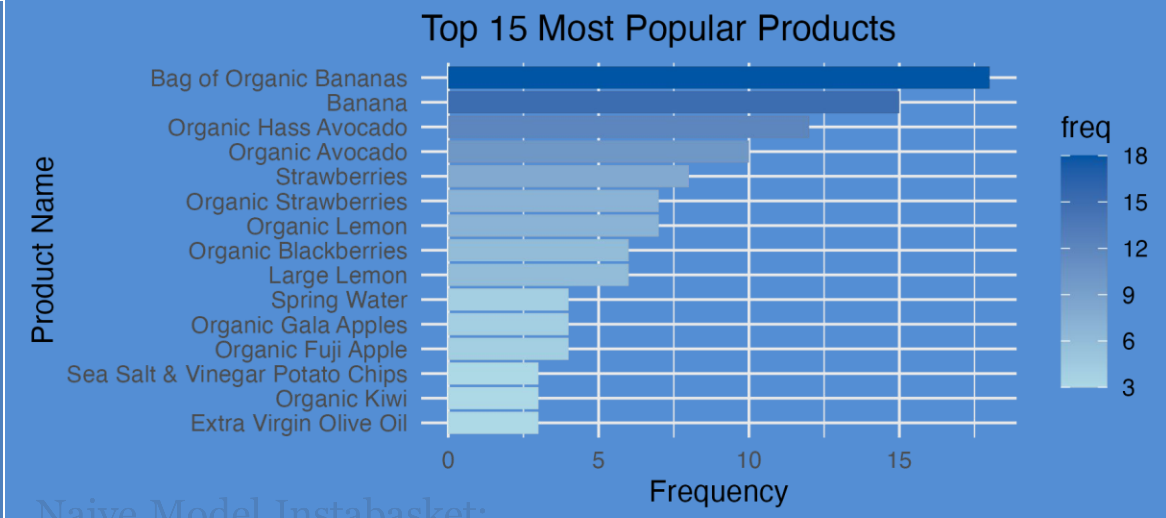
- 5% of the customer will take the 5% discount
- And will choose the Instabasket product



Rotman

Naive Model : *We consider the simplest model*

A straightforward way to select 1000 products in instabasket would be selecting **the most popular 1000 products with constraints**, which we called, the ***Naive Model***.



Naive Model Instabasket:

1	Banana	6	Organic Avocado
2	Bag of Organic Bananas	7	Large Lemon
3	Organic Strawberries	8	Strawberries
4	Organic Baby Spinach	…	…
5	Organic Hass Avocado	1000	Classic Soda Mini Bottles

An illustration from results

Check if the *constraints* are met:

> Yes, the result is as follows:

Frozen	Refrigerated	Other
100	100	800

Constraints: There are at most 100 products for each of frozen and refrigerated types

Result

	Metrics A	Metrics B
Training Set	56.5K/63K	46%
Testing Set	14.1K/16K	45%
Testing Set (With Substitutes)	14.3K/16K	48.7%

Metrics A: number of orders that utilize the in-aisle items

Metrics B: average % of items in each order that utilize in-aisle items

Comment

The Naive Model is valid and the 5% substitutes are have reasonably improve the metric!

Is this optimal?

The Naive model is tailored to optimize for popularity. However, our ultimate objective is to optimize for metrics A and B. At this juncture, we are uncertain if these two metrics have been effectively optimized.

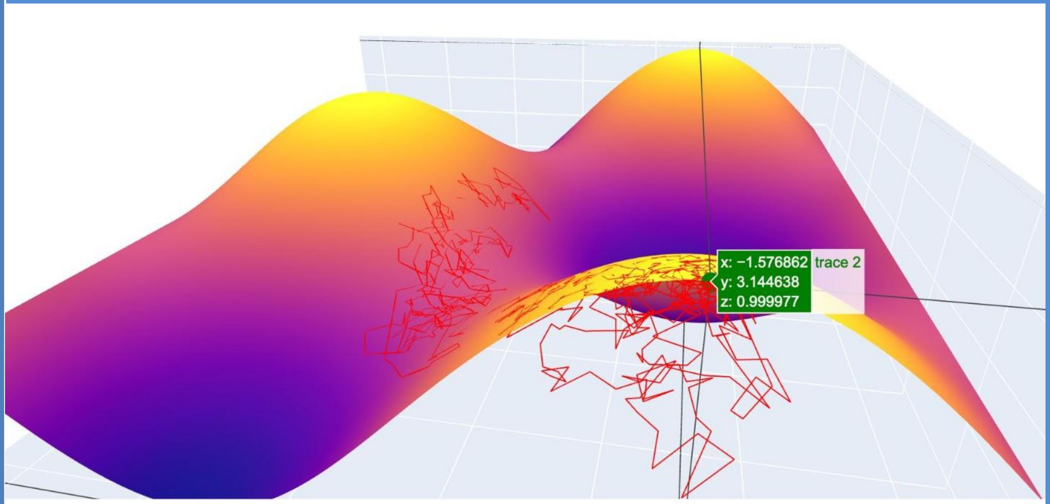
Second Model

Simulated Annealing Algorithm (Next Page)

Simulated Annealing Algorithm

Naive Algorithm is very straightforward… Can we try to find a better algorithm or prove the Naive Algorithm?

You can think of the *Simulated Annealing algorithm* as an explorer trying to find the lowest valley in a range of mountains. Initially, he might jump around trying to find a good starting point. But as time goes on, he becomes more cautious, focusing more on finding the true lowest point.



How it Works

1. Starting Point: We begin with a starting point, which can be any combination of products, for instance, the top 1000 products we've previously analyzed.
2. Tweak and Evaluate: We make a small change to this combination and then see if the new combination is better. If the new combination is better (i.e., more orders use these products), we accept it. If the new combination is not as good as our current one, we might still accept it sometimes, depending on factors like our current "temperature".
3. Role of Temperature: In the beginning, we set the temperature high, meaning we're more likely to accept solutions that aren't as good. This allows us to "jump around" in the solution space and not get stuck. But as time goes on, we reduce the temperature, making us more focused on finding the best solution.
4. Repeat: We repeat this process until the temperature is low enough or we've reached a maximum number of tries.

Result

In the end, we get a combination of products that perform the best out of all the combinations we've tried.

the number of orders that utilize the in-aisle items

	Metrics A	Metrics B
Training Set	56.4K/63K	46%
Testing Set	14.1K/16K	45%
Testing Set (With Substitutes)	14.3K/16K	49.1%

the average % of items in each order that utilize in-aisle items, accounting for any identified substitutes

Comment

The Simulated Annealing Model converges to the Naive Model

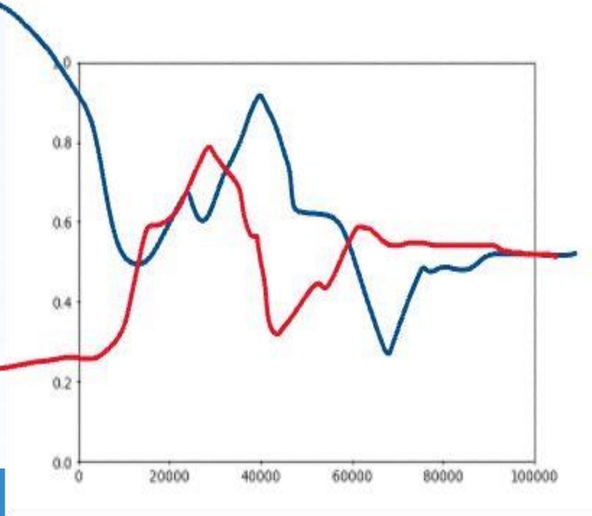
Result Analysis: Why are the results of the two algorithms surprisingly similar?

Naive Result

	Metrics A	Metrics B
Training Set	56.5K/63K	46%
Testing Set	14.1K/16K	45%
Testing Set (With Substitutes)	14.3K/16K	48.7%

Simulated Annealing

	Metrics A	Metrics B
Training Set	56.4K/63K	46%
Testing Set	14.1K/16K	45%
Testing Set (With Substitutes)	14.3K/16K	49.1%



100K Data Size

As the dataset grows larger, two different algorithms may converge to the same result if they are both designed to solve the same problem and are based on similar principles or assumptions.

As the dataset grows, our estimates become closer to the optimal value because we have more data to support our estimates. As a result, the two estimates will tend towards the same value, which is the ideal value.

Final Result:

Optimal Metric A: 14.3K/16K

the number of orders that utilize the in-aisle items is 14.3K out of 16K, nearly 89%

Optimal Metric B: 49.1%

the average % of items in each order that utilize in-aisle items, accounting for any identified substitutes, is 49.1%

Business Insight

Metrics A	Metrics B
14.3K/16K=89%	49.1%

Nearly 89% of orders would utilize the in-aisle items.

The average % of items in each order that utilize in-aisle items, accounting for any identified substitutes, is 49.1%

89% of orders are improved to reduce the time shoppers spend outside the 'Instabasket' and decrease the disruption caused by shoppers.

On average, for each order, 49.1% of the items can be found in the 'Instabasket'. This reduces the time shoppers spend outside of the 'Instabasket' by 49.1% for each order.

Using the Instabasket, we can:
1. *Improve efficiency* **of online order**
2. *Increase both satisfactions* **of online/in-store customers**

Limitation

The Instabasket doesn't fully account for the seasonal sales trends of products. Some products have a strong seasonal correlation, such as ice cream for summer and heaters for winter.
Our models aren't equipped to quantify the total time reduction for shoppers' visits to the supermarket
Our assumptions regarding substitutes could be enhanced with access to data on substitute records.

Future Directions (Beyond the Scope of This Study)

A/B Testing: After implementing the Instabasket strategy, we can conduct a survey on customer satisfaction. If satisfaction levels don't rise, we may need to consider further strategies, such as expanding the number of items in Instabasket from 1,000 to 2,000, or employing traversal algorithms to optimize the shopper's route outside of Instabasket.

Enhance the product counting system to accurately reflect the remaining quantity of items in the online app. Additionally, optimize the supply chain and establish a predictive model based on restocking patterns to forecast the frequency and quantity of replenishments. This ensures that shelves are consistently stocked, preventing situations where customers face out-of-stock items after making a purchase.

APPENDIX

Photo used in slide:
[1]Walmart. (2023). Retrieved from https://www.walmart.ca/en

Codes & Dataset used in the slide:
[2] Kuma110011. (n.d.). MMA TeamE. GitHub. Retrieved Oct 5, 2023, from https://github.com/Kuma110011/MMA-com-2023