Demand Forecasting with A.I.: Building the Business Case

By Neil C. Thompson, Nicholas J. Borge, Aparna Pande, Martin Fleming

1. Introduction

It is the second week of August 2018 and Mike Haydock, IBM's Chief Scientist, is preparing for a meeting with his supermarket client, CEO Lynne Marie of EuroGrocer supermarkets.[1] Haydock's group at IBM is advising France-based EuroGrocer on where they should apply machine learning within their supply chain and today Haydock will present the results of an initial pilot with his recommendations for roll-out.

IBM is coming off a success in applying machine learning to supermarket supply chains, having recently completed a 3-year project with large U.S. based grocer that also has a prominent position in international markets. That project showed that machine learning could improve the accuracy of demand predictions, helping the company ensure that the right number of products were on shelves to meet customer demand. The supermarket implemented IBM's recommendations across all stores in two product categories, and improved profitability by billions of dollars.

IBM is hoping to repeat their success in applying machine learning with EuroGrocer. Knowing that EuroGrocer's inventory has been growing significantly faster than revenues, Mike is sure that there is an opportunity to improve inventory management. However, he also knows that he can't simply copy what was done at previous supermarkets and assume that it will be successful. EuroGrocer faces its own challenges and would require its own tailored approach.

Despite being an important regional supermarket in Europe, EuroGrocer is much smaller than Mike's prior client. It also has less machine learning expertise and a different product mix, so the techniques that worked well before might be less promising at EuroGrocer. To understand how important these differences are, Haydock's team recently ran a pilot for EuroGrocer with two product sub-categories: *Ground Meat*, and *Yogurt*. Today's meeting is to report how successful the pilot was and make recommendations to Marie about where machine learning should be deployed.

As Haydock prepares for his meeting with Marie, he is confident that his team can pull off the technical part of the project. The key question is on the business side: where does applying machine learning make business sense? Haydock knows that answering that question means digging into the numbers.

[1] Note that the company name and some of the details of the case have been changed to be disguised.

2. Artificial Intelligence Primer

Artificial Intelligence (AI) is the ability of a computer to perform tasks commonly associated with intelligent beings, such as the ability to reason, discover meaning, generalize, or learn from past experience.[2] Early AI systems were rule-based; they attempted to reason about the world from a set of known facts. These systems were sometimes successful, but today have largely been superseded by systems that do statistical or "machine learning".

Machine learning is a branch of Artificial Intelligence focused on building applications that learn from data to identify patterns and make predictions. Instead of being trained with rules, they learn their rules from data, and thus are 'trained' to find patterns and use those patterns for predictions. Typical uses for machine learning include:

- **Clustering**: finding natural groupings in data based on similarities. For example, given a set of customers and their attributes (such as demographics and purchasing habits), segmenting them into like groups for marketing purposes.
- **Classification:** identifying which *class* (i.e., category) an observation belongs to, for example; given an image of an animal, determining which species it belongs to; given an email, determining whether it "spam" or "non-spam"; or given a customer service query, determining which product it is referring to. Note that in each of these examples, the classes (e.g., the list of species of interest) must be specified beforehand.
- **Prediction:** which covers many use cases, including predicting future values for a time series based on past buying trends and conditions. Predicting future customer demand for products in a supermarket is one example.

In recent years, machine learning has been successful in many fields. For example, machine learning is used to help interpret medical imaging, such as MRIs (Magnetic Resonance Imaging), and to understand human speech. machine learning helps Netflix to recommend shows to users, and Facebook to recognize faces in digital photos. Much of this success has come from moving from traditional machine learning techniques to *deep learning*.

2.1. Deep learning history

Artificial Neural Networks (ANNs, or simply NNs) mimic the human brain through interconnected neurons, or nodes (i.e., units of calculation), and were first proposed by Warren McCullough and Walter Pitts in 1943.[3] The first modern neural network was created by Frank Rosenblatt in 1953 with his so-called *Perceptron*, which consisted of one layer of neurons and was able to handle simple (linear) classification problems.[4] Despite the promise of NNs being

[2] https://www.britannica.com/technology/artificial-intelligence

[3] https://home.csulb.edu/~cwallis/382/readings/482/mccolloch.logical.calculus.ideas.1943.pdf

[4] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.335.3398&rep=rep1&type=pdf

able to solve more complex problems, computers in that era didn't have enough computing power to build NNs that were large enough to be useful, and so the field languished for decades.

Successive improvements in computer hardware made computers much more powerful, driven by Moore's Law.[5] This extra power unlocked the potential for *deep learning*, a technique that relies on *NNs* with more than three layers of nodes in depth. Deep learning models, with their greater number of interconnections between nodes, can model nonlinear relationships and solve more complex problems. Combined with the discovery that neural networks can be run much more efficiently on graphics processing units (GPUs), this meant that by the early 2010s computers were powerful enough to run even bigger, "*deeper*" networks.

In 2012 a famous deep learning model called "Alexnet" won an important image recognition contest, outperforming other algorithms.[6] Shortly thereafter, deep learning went on to break performance records in object detection, language translation, and speech recognition, demonstrating its amazing potential.[7,8,9] The corporate world has taken note of these successes and has invested heavily in recruiting, as has the finance world with massive increases in private investment in AI companies - see Exhibit 1.

2.2. Deep learning in practice

While the results that deep learning can achieve are impressive, many of the calculations are remarkably simple, usually just multiplication and addition. Combining them is what makes these systems powerful. For example, Figure 1 shows a typical graph of a deep learning system. Here the *nodes* indicate where calculations are performed, and the arrows indicate the flow of results from earlier calculations in the network to be included in later calculations. That is, the network illustrates the ordering for calculations that need to be done.

[5] Moore's law states that the number of transistors in a dense integrated circuit doubles about every two years.

[6] https://www.mygreatlearning.com/blog/alexnet-the-first-cnn-to-win-image-net/

[7] Historical progression of object detection algorithms: https://arxiv.org/pdf/1905.05055.pdf

[8] Google Neural Machine Translation upgrade to Neural Nets; https://ai.googleblog.com/2016/11/zero-shot-translation-with-googles.html

[9] https://towardsdatascience.com/a-brief-history-of-asr-automatic-speech-recognition-95de6c014187

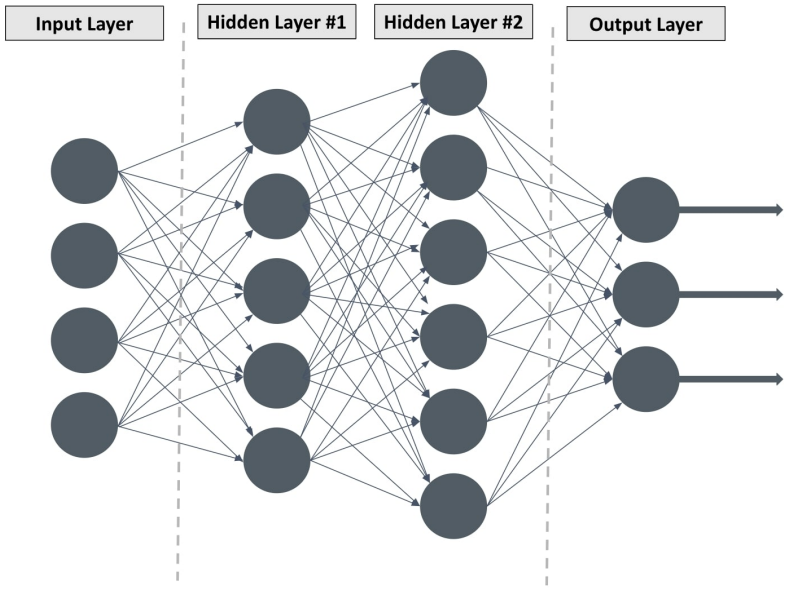

Figure 1: Illustration of a canonical neural network architecture, showing hierarchical layers of interconnected nodes

While the calculations at each node may be simple, an enormous number of them are often required. For example, let's take *EfficientNet,* a family of convolutional neural networks designed to reduce the number of calculations.[10] The EfficientNetL2 version needs a total of 51,840,000,000,000 calculations for a single training run which, on a cloud based processor running at 100 petaflops, takes a full 6 days to train. [11,12] This computation can be enormously expensive; two famous examples include $4 million to train OpenAI's *GPT-3* language model[13] and a whopping $35 million to train Deepmind's *AlphaGo Zero* to play *Starcraft 2*, a popular video game at the time.[14] As such, before a deep learning system is built, it is important that cost-benefit analyses are done to determine whether the system makes economic sense.

[10] https://arxiv.org/abs/1905.11946

[11] Petaflop = one thousand million million (10^{15}) floating-point operations per second

[12] https://arxiv.org/pdf/1911.04252v4.pdf

[13] https://spectrum.ieee.org/open-ais-powerful-text-generating-tool-is-ready-for-business

[14] https://www.yuzeh.com/data/agz-cost.html

3. Demand Forecasting

Demand forecasting is an important operational capability for firms. It tells them what customers are likely to want and thus allows the firm's supply chain and operational teams to have the right products and services available. Demand forecasts are used throughout both product and service-based businesses and are often the foundational assumptions for analyses of sales, profitability, cash flow, capital expenditure, and capacity planning.

The risk of not having enough of a product available to meet customer demand can be mitigated by holding lots of inventory, but this can be expensive. For supermarkets, whose profit margins are typically only 1-5%, being efficient with inventory is imperative.[15] This problem is even more acute for goods with a short shelf life, such as fresh produce, because of spoilage. If demand were constant over time, forecasting would be easy, but supermarkets must account for a range of complicating factors including seasonality, food trends, promotions, and weather-dependent buying.

Traditionally, stores depended on their managers' intuition and experience to decide on inventory. Managers would typically set fixed order amounts and then adjust them based on expected or observed changes in demand. Since 1982, demand forecasting and supply chain improvements have dramatically improved through better information technology, lowering inventory costs by 60% and transportation costs by 20%.[16]

The two most important reasons for managing supermarket demand well are as follows.

3.1. Avoiding Stock Outs

Stock-outs, where shoppers find that an item is not available for sale at a particular location, cost retailers an estimated $1 trillion every year worldwide.[17] Much of that revenue loss is gone forever, not just pushed to the next visit. In the longer term, repeated stock outs can diminish customer confidence and retention, creating opportunities for competitors. Though stock-outs can occur for a variety of reasons, some of which can be outside of the supermarket's control (e.g., global supply chain issues or unpredictable weather events), 72% of stock-outs are due to

[15] https://csimarket.com/Industry/industry_Profitability_Ratios.php?ind=1305

[16] Wilson, R. (2004). 15th Annual State of Logistics Report. Council of Supply Chain Management Professionals. Available online: http://www.cscmp.org/

[17] https://www.ihlservices.com/wp-content/uploads/2015/06/WeLostAustralia-Outline.pdf

inaccurate demand forecasts.[18,19,20] A recent study by the Grocery Manufacturers of America estimated that the average out-of-stock rate in the US is approximately 8%, though this varies significantly by product category.[21]

3.2. Avoiding Excess Inventory

73% of retailers struggle with forecasting inventory.[22] For some businesses it is easier to predict demand because it is highly seasonal (e.g., winter sports equipment stores). For grocery stores it is harder, because orders must often be placed weeks ahead of time and must consider a range of important factors such as last mile delivery logistics, customs, special storage requirements, etc. Supermarkets typically carry "safety stock", which acts as a cushion against unexpected surges in demand, but this also ties up additional working capital.[23]

The most substantial excess inventory cost faced by supermarkets comes from items that spoil or can't be sold. According to a 2011 CES Consumer Expenditures study, perishables account for 53% of US Supermarket sales.[24] Perishables that are unsold lead to "dead stock," like brown bananas, that need to be written off. The inventory lost to dead stock is particularly high in those product categories with shorter shelf lives, including meat, poultry, seafood, vegetables, and fresh fruit, at a rate of 4.5%, 9.7% and 11.4% of inventory respectively.[25] Discarded inventory will typically also have incurred "carrying costs" (e.g., electricity, warehousing, labor, rent, tax, and transportation costs), which are incurred whether or not these products are sold. Such costs are included in COGS (Cost of Goods Sold) and generally run between 20 percent and 30 percent of the total cost of inventory.[26]

3.3. EuroGrocer's Inventory Management

EuroGrocer is the second largest supermarket chain in France with 500 stores, $8B of annual revenue and a gross margin of 34.2%. Its retail business is highly seasonal so like most retailers it has a high turnover of store employees and hires staff to cope with seasonal demand. Every

[18] For instance, in Chile the number of products available in stores fell by 32% in the first two months after an earthquake https://www.nber.org/system/files/working_papers/w19474/w19474.pdf

[19] https://hbr.org/2004/05/stock-outs-cause-walkouts

[20] https://www.bostonglobe.com/2022/01/14/metro/why-so-many-empty-shelves-your-local-supermarket-its-complicated/

[21] https://www.nacds.org/pdfs/membership/out_of_stock.pdf

[22] https://www.globenewswire.com/news-release/2019/02/28/1744537/0/en/Automation-is-the-Key-to-Retail-s-Future-Survey-Says.html

[23] Weighted average cost of capital (WACC) for supermarkets is typically around 7.34%, according to https://finbox.com/NASDAQGS:SFM/explorer/wacc

[24] https://www.fmi.org/docs/facts-figures/grocerydept.pdf?sfvrsn=2

[25] https://www.ers.usda.gov/webdocs/publications/44100/eib-155.pdf?v=8907.6

[26] https://www.investopedia.com/terms/c/carryingcostofinventory.asp

one of EuroGrocer's stores has an automated store ordering system to predict and order the amount of goods that need to be in stock on any given day. The frequency of stock reordering varies depending on the product. Highly perishable products might be ordered daily, perhaps directly from local producers. High volume, low-value products may be ordered in bulk for cost efficiencies, either from warehouses or suppliers, and are thus re-ordered less frequently.

EuroGrocer's demand prediction system relies on certain basic optimizations based on product type and historical demand ranges to estimate future demand. It achieves a total stock loss of roughly 5.8%, of which 50% is due to spoilage, and a stockout rate of 3.4%. More data on how this breaks down by category is shown in Exhibits 3a and 3b, which also show the prediction error for each.

4. Economics of Artificial Intelligence

In the past decade, AI has seen major transformations, and substantially increased its impact on the world. Companies are hiring AI specialized engineers by the droves, and there has been a surge in investments. AI startups raised $73.4B in cumulative funding in 2020 (up from $5.5B in 2015), and government interest in AI has also grown (e.g., the US government now spends billions of dollars across multiple programs).[27]

Though the applications of AI are vast, companies are limited by cost and capacity constraints as they implement machine learning solutions. The three biggest cost areas involved in developing and maintaining a machine learning project within a company are labor (personnel), data, and computation costs.

Labor costs cover the software engineers, data scientists, and subject matter experts used to develop and maintain the AI systems, which can include both full time employees and contractors or consultants. For EuroGrocer, each full-time employee in this category costs $200,000 per year (including salaries, bonus, benefits) on average. For temporary staff, EuroGrocer commonly pays $40,000 per month per engineer on average. These labor costs typically fall into 2 phases of a machine learning project: (1) design and implementation, and (2) on-going maintenance.

Data costs for a machine learning project can be extensive. A simple machine learning model might require only up to 1000 data points to train, but more complex models can require millions of data points to make their results broadly applicable and mitigate algorithmic bias.[28] The cost of obtaining data for an AI can vary from purchased data (which can be quite expensive), to freely available data which may or may not require labeling by subject matter

[27] https://stockapps.com/blog/2020/11/18/ai-startups-raised-73-4b-in-total-funding-over-15b-of-investments-in-2020/

[28] https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

experts. Data used to train and test models must also be stored for when re-training is required. EuroGrocer pays $0.01 per GB (gigabyte) per month for on-going storage costs.

Compute costs for running neural networks can be significant, because of the large numbers of calculations required. Principally this consists of:

1. **Training**: Iteratively analyzing historical data or labeled examples to learn the "rules" for making successful predictions. Typically, this is done once up-front and then repeated periodically based on how quickly circumstances change and thus "new rules" need to be learned.
2. **Inference:** The active usage of the machine learning models to make predictions. This is done whenever new predictions are needed e.g., whenever stocks of goods need to be replenished.

The costs are accounted for differently based on whether hardware is purchased outright (i.e., a capital cost) or whether computing time is purchased from a cloud provider (i.e., a variable cost). EuroGrocer uses a cloud service and pays $7.894 / hour / instance for cloud compute.

5. EuroGrocer AI demand forecasting pilot

5.1. Setting up and running the pilot

A IBM team of 5 full time engineers and one program manager set up the system to make predictions for each SKU-store combination. For the pilot, two models were built, one for each of *Ground Meat* (a subcategory of the *Meat and Seafood* product category) and *Yogurt* (a subcategory of the *Dairy, Eggs and Cheese* product category). Together, these two models covered 15,000 SKU-store combinations (500 stores x ~30 SKUs per store) during the pilot.

The team worked on the pilot for 6 months (from January 2018 to May 2018), alongside two EuroGrocer engineers (at 80% capacity), two Eurogrocer analysts (at 80% capacity), and with an additional combined 9 hours per month of support from EuroGrocer subject matter experts (e.g., store managers). Together, they tested a variety of machine learning approaches, looking at prediction error, computation complexity, equipment cost and ease of iteration.

They incorporated two years of historical data on sales (the "target" data that would also be used for calibration of the models), along with internal data from EuroGrocer's ERP (Enterprise Resource Planning) systems on product attributes, and information from customer loyalty programs. They also looked at a range of external data that were available through EuroGrocer's analytics function e.g., seasonal holidays, broader consumption patterns and macroeconomic factors.

It took several months of piloting the new system until it was stable and mature. At that point, the two models took 16 hours to train and 4 hours to run across the 15,000 SKU-store combinations, taking up 95 GB of storage.

The results showed a marked improvement in prediction accuracy. Using the new demand forecasting algorithm, the prediction error dropped from 32.5% to 18.7% for *Ground Meat* and from 31.0% to 17.9% for *Yogurt*. These outcomes did not quite reach the anticipated goal of 10% prediction error for but were considerable improvements, nonetheless. Thus, on average, prediction errors dropped 42.4%. Based on the pilot, the team determined that the models needed to be trained once every two weeks and run once every day to predict demand.

5.2. Implications for roll-out

Based on their experience with the pilot, the team estimated the effort required to implement the wider program and support ongoing efforts. They reasoned that the project would take a one-time effort of 6 months to set up, supported by the same IBM and EuroGrocer team from the pilot. The EuroGrocer team would be able to support and maintain the system on their own, at only 30% of their capacity, allowing them to work on other projects in parallel, and with only an estimated 2.5 hours per month from EuroGrocer subject matter experts.

Based on this experience, the team decided that the appropriate level of aggregation for models was at the product category, not subcategory, level. This means a model for each product category (e.g., *Dairy, Eggs, and Cheese*), that would be trained on data across all SKU-store combinations belonging to that category (i.e., all combinations in all subcategories such as *Yogurt*). Alternative options considered included doing one model per store, which could better focus on regional effects, but would yield too many models to maintain. Modeling by product category would yield up to 19 models (if EuroGrocer decided to proceed for all products), which the team felt would be a good balance between predictive power and ease of management.

There would be no additional costs for data collection or incremental infrastructure, as EuroGrocer's systems were already configured with appropriate APIs, and storage and compute would scale based on the number of SKU-store combinations.

6. Decision

As Mike reflected on the experience with the pilot, he knew that he would need to make a business case grounded in the economic benefits and costs of the proposed system, which he organized as follows:

Benefits of demand prediction using deep learning:

1. How valuable would it be to the business if it were possible to completely eliminate stock-outs arising from incorrect demand prediction? What about spoilage?
2. If deep learning provided a 1 percentage point improvement in prediction accuracy for each of EuroGrocer's product categories, how big an effect would it have on profitability and which categories would drive the most benefit? The least? Why?
3. If the accuracy improvement from the pilot is representative, what would be the total annual benefit from using deep learning across all product categories?

Cost of demand prediction using deep learning:

4. For each category, how much would it cost per year to run the deep learning system?

Implications of demand prediction using deep learning:

5. For each category, what is the net benefit from applying the deep learning system?
6. Discuss whether Mike should recommend to Lynne Marie that EuroGrocer adopt this system. For which categories? Is the upfront investment justified?
7. What other strategic or operational considerations should EuroGrocer consider before adopting the system?

Appendix

Exhibit 1: Private investment in Funded AI companies 2015-20. Source: CAPIQ, Crunchbase and NetBase Quid, in Stanford AI index)
https://aiindex.stanford.edu/wp-content/uploads/2021/11/2021-AI-Index-Report_Master.pdf

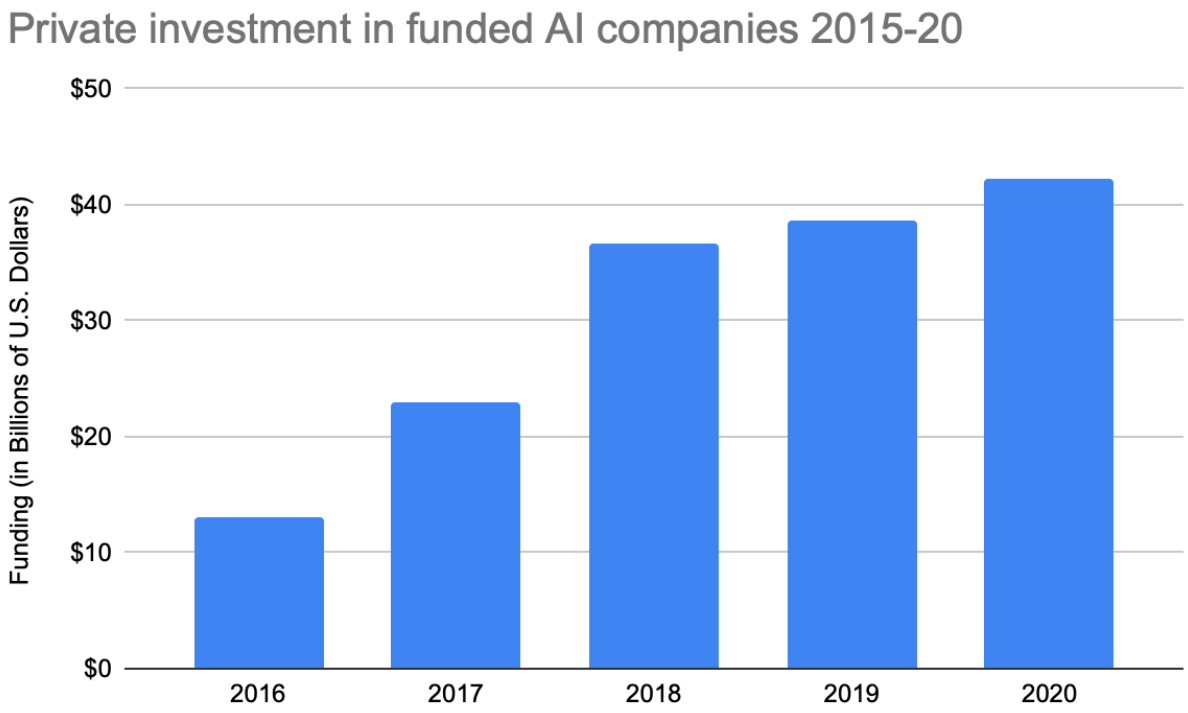

Exhibit 2: Cost Drivers for A.I. Models

This cost tree outlines some of the key components in calculating the total cost of a machine learning system.

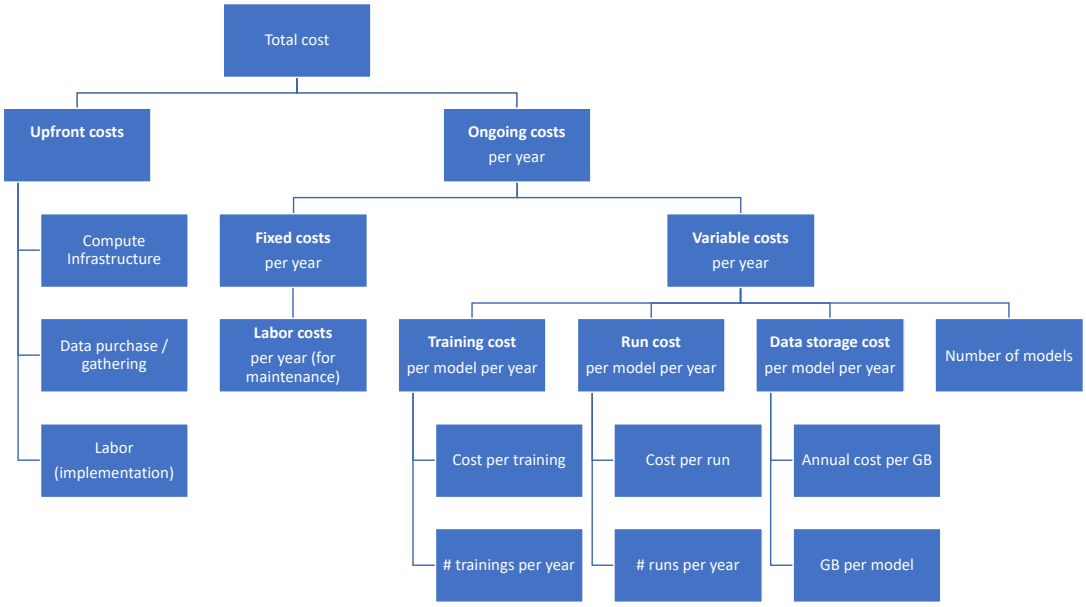

Exhibit 3a: Product category data (table 1 of 2)

Parent product category	Product category	Avg. # of SKUs per store	Annual revenue ($M)	Gross margin (%)	Average price bracket	% online Shoppers who buy this
Fresh	Bakery	1,490	184	55	Low	30
Fresh	Beers, Wine and Spirits	980	462	23	High	30
Fresh	Dairy, Eggs & Cheese	2,980	852	31	Low	24
Fresh	Deli	2,620	567	40	Medium	30
Fresh	Floral	150	47	50	Med	30
Fresh	Meat & Seafood	1,490	1,008	30	Medium	17
Fresh	Produce: Fruits & Vegetables	3,140	866	45	Low	22
Frozen foods	Frozen Foods	1,590	593	30	Medium	30
General merchandise	Cleaning Supplies	460	221	47	High	35
General merchandise	Paper Products	210	158	47	Low	42
General merchandise	Pet Care	510	95	47	High	28
General merchandise	Tobacco	210	32	25	High	15
Health and beauty	Baby	620	126	47	Medium	35
Health and beauty	Health and beauty	2,570	377	54	High	45
Shelf stable foods	Canned Goods & Soups	1,130	362	25	Low	37
Shelf stable foods	Condiments/Spices & Bake	3,600	241	20	Low	31
Shelf stable foods	Dry Goods, Shelf Stable Foods	4,630	482	25	Low	37
Shelf stable foods	Grains, Pasta & Sides	1,290	844	25	Low	30
Shelf stable foods	Salty Snacks	1,440	482	37	Low	43

Exhibit 3b: Product category data (table 2 of 2)

Parent product category	Product category	Current MAPE (%)[29]	Stockouts (% of revenue)	Average Shelf Life	Stock loss (as % of revenue)	% of stock loss that is due to perishability
Fresh	Bakery	31.3	6.0	1-3 weeks	2	60
Fresh	Beers, Wine and Spirits	30.3	2.3	Varies	5	15
Fresh	Dairy, Eggs & Cheese	31.6	2.1	1-14 days	7	60
Fresh	Deli	31.8	3.3	3-5 days	5	60
Fresh	Floral	30.4	3.0	7-12 days	1	70
Fresh	Meat & Seafood	40.0	7.8	2-5 days	10	70
Fresh	Produce: Fruits & Vegetables	38.3	4.7	Varies	12	80
Frozen foods	Frozen Foods	31.4	3.5	8-12 months	5	45
General merchandise	Cleaning Supplies	30.0	2.4	High	1	0
General merchandise	Paper Products	30.0	2.2	High	1	0
General merchandise	Pet Care	30.1	1.9	Varies	4	10
General merchandise	Tobacco	30.0	3.0	High	2	0
Health and beauty	Baby	30.1	1.9	3-9 months	2	10
Health and beauty	Health and beauty	30.1	1.9	2-3 years	6	5
Shelf stable foods	Canned Goods & Soups	30.1	1.8	2-3 years	1	15
Shelf stable foods	Condiments/Spices & Bake	30.1	1.5	1-2 years	2	10
Shelf stable foods	Dry Goods, Shelf Stable Foods	30.0	1.8	6-24 months	2	5
Shelf stable foods	Grains, Pasta & Sides	30.2	1.8	2-3 years	5	10
Shelf stable foods	Salty Snacks	30.4	3.8	High	5	10

[29] https://www.statisticshowto.com/mean-absolute-percentage-error-mape/