FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Machine Learning-based approaches for discount policies in food retailing

Sónia Ferreira

Mestrado em Engenharia e Ciência de Dados

Proponent: Sara Martins (INESC TEC advisor)

Supervisor: Maria João Santos

Supervisor: Pedro Amorim

October 28, 2024

Machine Learning-based approaches for discount policies in food retailing

Sónia Ferreira

Mestrado em Engenharia e Ciência de Dados

October 28, 2024

Abstract

The present research investigates how applying machine learning (ML) techniques can optimize discount policies in the retail sector, focusing on perishable products. Because perishable products represent a significant part of retail inventory and are inherently time-sensitive, they pose a continuous challenge for retailers trying to reduce losses while maintaining profitability. Therefore, this research aims to address the challenge of maximizing profitability for these products by exploring more efficient pricing strategies.

Dynamic pricing, a strategy that involves real-time price adjustments based on factors like demand, inventory levels, and market conditions, has shown significant potential in enhancing profitability in the retail industry. This potential is particularly promising for perishable products, where dynamic pricing models can optimize discounting strategies, ensuring that these goods are sold before they spoil. This reduces financial losses from unsold inventory and boosts overall revenue.

The review of existing literature outlines several gaps in current dynamic pricing strategies, particularly the need for empirical validation and adaptation to the unique characteristics of different types of perishable products. Additionally, it highlights the importance of integrating consumer behavior insights into pricing models to better align discount policies with customer demand and purchasing behavior. By employing reinforcement learning (RL) techniques, the research aims to optimize discount policies specifically for perishable products, with the dual objective of enhancing profitability and reducing waste. While reducing food waste is desirable, this research focuses on how these techniques can improve profit growth.

The results show that applying RL algorithms to dynamic pricing models can significantly improve profitability for perishable products. Among the algorithms tested, Trust Region Policy Optimization (TRPO) emerged as the most effective at maximizing profitability, delivering the highest overall profit. However, this came at the cost of increased waste, highlighting the trade-off between aggressive discounting strategies and waste management. While TRPO leads in profit maximization, retailers must carefully manage the resulting waste to ensure long-term sustainability. Finally, an optimal discount range of 30%-60% was identified, encouraging consumer purchases while balancing sales and profitability.

Keywords: Machine Learning, Reinforcement Learning, Dynamic Pricing, Retail, Food Retail Sector, Consumer Behavior, Discount Strategies, Food Waste, Perishable, Economic Efficiency

Resumo

A presente investigação, explora como a aplicação de técnicas de machine learning (ML), pode otimizar a aplicação de políticas de desconto no setor retalhista, com foco particular nos produtos perecíveis. Uma vez que os produtos perecíveis representam uma parte significativa do inventário no setor de retalho e são intrinsecamente sensíveis ao tempo, estes constituem um desafio constante para os retalhistas, que tentam reduzir perdas, mantendo, ao mesmo tempo, a rentabilidade. Portanto, esta investigação pretende abordar o desafio de maximizar a rentabilidade destes produtos, explorando estratégias de preços mais eficientes.

Os preços dinâmicos são uma estratégia que envolve o ajuste de preços em tempo real, com base em fatores como a procura, os níveis de inventário e as condições de mercado. E tem demonstrando um potencial significativo para aumentar a rentabilidade no setor do retalho. Este potencial é particularmente promissor para produtos perecíveis, onde os modelos de preços dinâmicos podem otimizar as estratégias de descontos, garantindo que estes bens são vendidos antes de expirarem. Isto reduz as perdas financeiras resultantes do inventário não vendido e aumenta assim as receitas totais.

A revisão da literatura existente destacou várias falhas nas estratégias de preços dinâmicos atuais, em particular a necessidade de validação empírica e adaptação às características únicas de diferentes tipos de produtos perecíveis. Além disso, sublinha a importância de integrar insights sobre o comportamento do consumidor nos modelos de preços, de forma a alinhar melhor as políticas de descontos com a procura e o comportamento de compra dos clientes. Utilizando técnicas de reinforcement learning (RL), esta investigação tem como objetivo otimizar as políticas de descontos especificamente para produtos perecíveis, com o duplo objetivo de aumentar a rentabilidade e reduzir o desperdício. Embora a redução do desperdício alimentar seja desejável, o foco da investigação está em como estas técnicas podem melhorar o crescimento dos lucros.

Os resultados mostram que a aplicação de algoritmos RL em modelos de preços dinâmicos pode melhorar significativamente a rentabilidade de produtos perecíveis. Entre os algoritmos testados, o Trust Region Policy Optimization (TRPO) destacou-se como o mais eficaz na maximização da rentabilidade, obtendo o maior lucro global. Contudo, isto teve como custo um aumento de desperdício, evidenciando o compromisso entre estratégias de desconto agressivas e a gestão de desperdícios. Embora o TRPO seja o melhor em maximizar os lucros, os retalhistas devem gerir cuidadosamente o desperdício resultante para garantir a sustentabilidade a longo prazo. Por fim, foi identificada uma faixa de desconto ideal de 30% à 60%, que deverá incentivar às compras dos consumidores e ao mesmo tempo equilibrar às vendas e a rentabilidade.

Keywords: Machine Learning, Reinforcement Learning, Preços dinâmicos, Retalho, Retalho alimentar, Comportamento do consumidor, Estratégias de desconto, Desperdício alimentar, Perecíveis, Eficiência económica

Acknowledgements

To all my friends for their patience when I was not present on many occasions and for always being present in my life.

To my colleagues who helped me achieve my academic goals by providing a good environment, working together, and being good teachers when I needed them.

Finally, I want to thank my supervisors and the team for their support throughout this process and for their willingness to help me and give suggestions to improve my work.

Sónia Ferreira

"Do the best you can
until you know better.
Then when you know better,
do better."

- Maya Angelou

Contents

List of Figures

Abbreviations

A2C	Advantage Actor-Critic
BOGOF	Buy one, get one free
EE	Energy Efficiency
FAO	Food and Agriculture Organization
FL	Food Loss
FLW	Food Loss and Waste
FW	Food Waste
MDP	Markov Decision Process
ML	Machine Learning
MLOR	Minimum Level of Receipt
PPO	Policy Optimization
RL	Reinforcement Learning
R-PPO	Recurrent Proximal Policy Optimization
SDG	Sustainable Development Goal
TRPO	Trust Region Policy Optimization

Chapter 1

Introduction

In the past years, there has been a concerning increase in the number of people affected by extreme hunger. Global hunger has remained relatively unchanged from 2021 to 2022 but remains alarmingly above pre-COVID-19 pandemic levels, affecting around 9.2 percent of the world population in 2022 compared with 7.9 percent in 2019 (FAO, 2023). This equates to an estimated 691-783 million people facing hunger in 2022, underscoring a persistent global challenge (Food Security Information Network, 2023). Consequently, this puts us further away from achieving the first target of the second Sustainable Development Goal (SDG), defined in 2015, which was jointly agreed upon by the United leaders: "By 2030, end hunger and ensure access by all people, in particular, the poor and people in vulnerable situations, including infants, to safe, nutritious and sufficient food all year round" (United Nations, 2024).

The seriousness of acute food insecurity is primarily attributed to conflict, economic turmoil, and extreme weather, all of which have become increasingly prevalent globally (Food Security Information Network, 2023). Additionally, an enormous amount of food is lost and wasted daily, exacerbating this issue. Food loss primarily happens in the initial phases of the food supply chain, including production, harvesting, and processing, often due to inefficiencies and lack of infrastructure. On the other hand, food waste predominantly occurs at the consumer level in households, and retail is frequently linked to over-purchasing, poor planning, and strict aesthetic standards for produce.

According to ReFED (2024), there are three main sources of food waste: households, food services, and retailers. Even though retailers rank third in the global waste hierarchy, they are responsible for 13% of food waste along the supply chain (United Nations Environment, 2021). As they bridge the gap between food production and consumption, their contribution to reducing food waste can be crucial. By optimizing food distribution and pricing strategies, particularly for perishable products (in particular for the ones with a nearing expiration date, which account for about fifty percent of retailers' sales (Li et al., 2012)), retailers can significantly reduce food waste, optimize their profitability, ensure better food distribution, and potentially alleviate some aspects of global hunger.

Perishable products play a key role in the dynamics of food waste, a role that is significantly influenced by whether or not they bear expiration dates. Perishables lacking a defined expiration date pose particular challenges. These challenges stem largely from the variability in consumer demand and the rigorous aesthetic standards that retailers and consumers enforce. Such standards often lead to discarding food that fails to meet these subjective criteria, regardless of its actual edibility or nutritional value. Conversely, perishable goods featuring expiration dates have a different impact on food waste. Consumer misunderstanding of labeling terms like "sell by" and "use by" frequently results in the unnecessary disposal of still-edible food. Consumer purchasing behaviors exacerbate this phenomenon, as noted by Harcar and Karakaya (2005), most consumers prioritize products with longer remaining shelf lives in their purchasing decisions. This preference is further corroborated by Konuk (2018), who emphasizes the inclination towards perishables with a more extended shelf life.

It's important to acknowledge the efforts already made by retailers to address the issue of perishable products with short shelf lives. As Chen et al. (2016) highlighted, numerous retailers have undertaken initiatives to prevent such products from lingering on shelves until the point of mandatory removal. Specifically, they began by introducing discounts to compensate for the perceived lower utility of these products (Chung, 2019). The strategy outlined by Chung and Li (2014) reflects a practical approach retailers adopt to manage perishable products nearing the end of their shelf life. This approach involves applying discounts ranging from 20% to 50% during the last days of the perishables' shelf life.

Even though some studies have already focused on supporting and updating depreciation policies by investigating models capable of optimizing dynamic pricing (Tsiros and Heilman, 2005), these approaches exhibit notable gaps, particularly in their assumptions about consumer behavior. A persistent issue is using pre-defined, theoretically constructed functions to predict consumer demand, often lacking empirical validation. This methodological shortfall has led to inadequacies in price discrimination throughout the shelf life of food products. Additionally, the failure to tailor discounts to different products' unique characteristics has resulted in skewed sales forecasts and inaccuracies in strategies aimed at maximizing profits and minimizing waste, as Popescu and Wu (2007) noted.

These challenges highlight the need for more nuanced and empirically grounded models to address the complexities of perishable goods management. Implementing optimal discount pricing strategies for perishable goods nearing expiration is a potential solution to curb food waste. Yet, this approach does not come without its challenges. Retailers are faced with the dual task of devising discount strategies that appeal to consumers while ensuring that these reduced prices do not negatively impact their revenue.

1.1 Context and Motivation

The significance of this research stems from the growing concern about food waste and its environmental, economic, and social implications. These alarming statistics highlight the urgent need for effective and sustainable solutions within the food system, especially in the context of food retailing. Managing food resources, particularly in addressing waste and inefficiencies, becomes crucial to a broader strategy to combat hunger and food insecurity worldwide.

Hence, the research intended to be carried out seeks to support retailers, through data-driven solutions, in efficiently managing perishable goods with fixed expiration dates and mitigating food waste while maximizing profitability at the store level. It's driven by a larger initiative entitled *BeFresh: Integrating Consumer Behavious to Improve Food Value Chains*, which is aligned with three of the Sustainable Development Goals (SDGs), namely to eradicate hunger, promote inclusive and sustainable economic growth and ensure sustainable consumption and production patterns (INESC TEC, 2024). *BeFresh* project encompasses four main tasks:

- **Task 1:** Involves demystifying the relationship between consumer choices and the dichotomy of perishable goods' shelf life versus their pricing. It includes studying the effects of consumer preferences on product substitution, especially between items with varying prices and expiration dates.
- **Task 2:** The primary focus of this thesis research involves creating dynamic pricing policies for perishables throughout their shelf life spectrum. This aims to balance consumer demands with retailer needs, reducing the amount of food that expires unsold on shelves.
- **Task 3:** Involves analyzing and optimizing crucial parameters in the value chain, such as the Minimum Level of Return (MLOR) for groceries.
- **Task 4:** It will be attempted to create synergies between retailers and producers to reduce costs and food waste in the retailer's chain.

While developing effective discount strategies for perishables (task 2) using data-driven solutions is currently the main area of development of this research, its outcomes are expected to impact and inform the approaches to be taken in tasks 3 and 4 significantly.

The significance of this thesis is underscored by its practical application in collaboration with one of Portugal's largest retailers, which deals with food waste brought on by its products' perishability. The research will utilize historical data from this retailer's yogurt portfolio, including daily sales, inventory receipts, and in-store stock levels. This data-rich environment is ideal for applying and evaluating the proposed machine learning (ML) based approaches.

1.2 Research Objectives

In light of the aforementioned context, it becomes imperative to develop and integrate effective, sustainable ML-based approaches within the food system, with a specific emphasis on the domain of food retailing.

The retail food industry faces a critical challenge in managing perishable products, which are integral to both financial viability and consumer satisfaction. One of the most pressing issues is the optimal pricing of perishables to minimize food waste while maximizing profits, aiming to strike a harmonious balance between environmental sustainability and economic viability.

Against this backdrop, the following research question was raised:

- What is the optimal discount range that encourages consumers to purchase perishables products and maximize retailer profitability?

1.3 Contributions

As previously mentioned, strategically managing perishables is a difficult but crucial task in the retail industry. Thus, this research aims to contribute toward a more efficient, economically viable, and environmentally sustainable food retail system. The potential benefits of integrating ML techniques with practical retail concerns include a considerable reduction in food waste, a step towards alleviating hunger, and bolstering economic efficiency within the food retail sector.

- **Impact on Food Waste Reduction:** By optimizing the pricing and discount strategies for perishable goods, retailers can significantly reduce the amount of food that goes unsold and ultimately wasted. This approach helps manage inventory more effectively and contributes to a broader social cause – reducing hunger. Making perishable food items more affordable and accessible can be a key factor in feeding more people, especially those in lower-income groups.
- **Maximizing Retail Profits:** An effective pricing strategy ensures that perishable goods are sold before expiring, minimizing spoilage losses. This approach also aims to find a sweet spot where the price reductions, either due to frequent updates or discounts, do not significantly undercut the retailer's profit margins. The goal is to create a win-win scenario where food wastage is minimized, and profitability is maximized.

In conclusion, the strategic management of perishables within the retail sector holds great promise for individual businesses and society as a whole.

1.4 Document Structure

Besides the present introductory chapter, Chapter 2 delves into existing research related to food wastage, the role of the retail sector in contributing to and mitigating this issue, and the advancement of retail pricing strategies with a focus on perishables. This chapter synthesizes and exposes the current findings and identifies gaps that our research aims to fill.

Chapter 3 outlines the research methodology employed in the research. It details the algorithms and models used, the data collection process, and the experimental setup. In chapter 4, the results are presented and discussed. Chapter 5 concludes by giving a brief overview of the key findings, discussing their implications, and providing recommendations for future research while also reflecting on the limitations of the current research, and suggesting potential areas for further exploration.

Chapter 2

Literature Review

The current chapter will address the critical issue of perishable products management, specifically in the retail sector, highlighting the need for strategies that not only minimize waste but also maximize profitability. Building on these themes, the existing literature will be examined, focusing on retail strategies, consumer behavior, and the potential use of machine learning in frequent price adjustments. The goal is to investigate how these interconnected factors contribute to both the reduction of perishable product losses and the improvement of financial performance for retailers, highlighting the balance between reducing waste and enhancing profitability.

2.1 Introduction

In the preceding chapter, a review of several interlinked global challenges was discussed. Central to these is the alarming rise in global hunger, which has reemerged as a pressing issue. This is closely linked with the prevalent issue of food waste, particularly within the retail sector. A sector that is critical to the supply chain but suffers significant losses, particularly with perishable goods. These waste patterns exacerbate the hunger crisis and pose a challenge to sustainable resource management. Furthermore, the issue extends beyond environmental and social implications, as it directly impacts the economic viability of retailers, as significant food waste translates into substantial financial losses, underscoring the urgency to explore efficient management strategies for perishables. Against this backdrop, the necessity of approaches that balance profitability and sustainability becomes clear as retailers must navigate the delicate equilibrium between minimizing waste, fulfilling consumer needs, and maintaining economic sustainability. In this context, exploring effective management strategies, including technological innovations, is an ethical and environmental concern and a key economic consideration.

The objectives of the present literature review are multifaceted and aim to delve deeper into the complexities of these interconnected themes. Primarily, this chapter seeks to investigate the existing body of research in these areas. Most of the focus will be on the retail strategies currently being adopted or proposed to mitigate food waste. Understanding consumer behavior in this context is vital since it directly influences retailers' practices and waste patterns.

Furthermore, in an era where technology is increasingly used to shape business strategies, this review will look at the potential use of machine learning to facilitate a more frequent price adjustment. Unlike traditional dynamic pricing, which adjusts prices in real time based on various factors such as demand, supply, and shelf-life, this approach aims to redefine the frequency and regularity of price updates to assess its efficacy in reducing food waste. More systematic and optimal pricing update periods could be a critical factor in reducing food waste, thereby enhancing the economic health of retailers and contributing to larger goals, such as reducing global hunger and improving sustainability in the retail sector.

In resume, this chapter seeks to provide a comprehensive overview of the current literature on these domains, offering insights into current practices, challenges, and potential technological innovations that can revolutionize the retail sector's approach to managing perishables.

2.2 Food Wastage, Loss and Waste

Food wastage is an important global issue, marked by the loss and waste of valuable food resources. However, in academic literature, the definition of food wastage is not universally standardized and differs depending on the context. Despite numerous papers acknowledging this issue and several researchers discussing it, no single, widely accepted definition of food wastage remains. This lack of consensus extends to concepts like food loss and waste, highlighting the variability and complexity in scholarly interpretations of these terms.

In the academic article, "Can we agree on a food loss and waste definition? An assessment of definitional elements for a globally applicable framework," Boiteau and Pingali (2023) embark on a rigorous examination of the conceptualization of food loss and waste (FLW) in the pursuit of sustainable food systems, notably aligning with the first target (target 12.3) of the second Sustainable Development Goal (SDG). This critical analysis is positioned within the broader discourse on the standardization of FLW definitions and their implications for global sustainability initiatives.

The authors' methodological approach involves an extensive review and comparison of prevailing FLW definitions across a comprehensive dataset from the FAO FLW database, encompassing over 21,786 data points from 2004 to 2021. This dataset serves as a foundation for assessing the distribution and availability of FLW data across contrasting country income groups, thereby illuminating regional disparities and focal areas within the FLW narrative.

A key aspect of their study is the comparison of existing FLW definitions against the FAO's proposed 2014 standard. This comparison extends to the Food Loss Index and Food Waste Index, both integral sub-indicators for realizing SDG target 12.3. The study identifies and critiques discrepancies regarding utilization and edibility criteria, which must should be addressed more consistently across different definitional frameworks.

The authors advocate adopting the FAO's 2014 FLW framework, positing it as comprehensive and expansive for global application. This framework succinctly encapsulates FLW as the diminution in either quantity or quality of food intended for human consumption. This definition encapsulates scenarios ranging from the diversion of food to non-food applications to the degradation in nutritional value, safety, or other qualitative aspects from the point of harvest or slaughter to the point of consumption.

The challenge in defining food wastage lies in its scope and the stages at which it occurs. The ambiguity in distinguishing between food loss and food waste can lead to challenges in developing targeted strategies for mitigation. Moreover, different interpretations of what constitutes 'edible' food further complicate this definition. For instance, cultural and regional differences can influence perceptions of edibility and what is classified as waste.

Therefore, it is vital to distinguish between these terms to develop a clear understanding of where interventions can be adequate. For example, strategies to reduce food loss may focus on improving agricultural practices, enhancing storage and transportation, or advancing processing techniques. On the other hand, initiatives to curb food waste can target consumer education, changes in retail practices, or policies that promote food donation.

In this context, we will adopt the following definitions for food loss and waste, as provided by HLPE (2014):

- Food losses (FL): refers to a decrease at all stages of the food chain before the consumer level (from production to distribution), in mass, of food initially intended for human consumption, regardless of the cause.
- Food waste (FW): refers to food appropriate for human consumption being discarded or spoiled at the consumer level, regardless of the cause.

In other words, food loss typically refers to decreased quantity or quality at the production, post-harvest, and processing stages of the food supply chain. In contrast, food waste primarily occurs at the retail and consumption stages and is often linked to retailer and consumer behavior. Therefore, food waste entails the removal of food that is fit for consumption, either due to deliberate choices or negligence leading to spoilage. Although food waste is part of food loss, it differs from unintentional food loss in its underlying reasons and motivations (Boiteau and Pingali, 2023).

2.3 Food Waste in the Retail Sector

In the global waste hierarchy, food waste in the retail sector is significant, ranking third in its impact and magnitude (ReFED, 2024). Retailers play a pivotal role as key intermediaries in the food supply chain. Factors contributing to food waste include inefficient supply chain strategies, lack of preventive approaches, and a focus on market demand rather than the shelf life and loss characteristics of products (Muriana, 2017). The consequences of this food waste are extensive, affecting not only the economic bottom line of retailers but also contributing to environmental issues such as resource scarcity and increased greenhouse gas emissions.

Managing food waste in the retail sector is an ethical and practical need to ensure food security, conserve resources, and mitigate environmental effects. As central players in the supply chain, retailers can reduce waste significantly. They can influence both upstream (production) and downstream (consumption) waste through a combination of policy intervention, technological innovation, and behavioral change across all supply chain levels (Teller et al., 2018). Namely, they can include strategies to optimize supply chain logistics, invest in waste-reducing technologies, and engage in consumer education initiatives.

2.4 Retailers' Strategies to Manage Food Waste

Retailers need help managing food waste, particularly in the realm of perishable products. Yet, managing food waste presents a complex and multifaceted challenge as they face several critical issues. Namely, maintaining optimal inventory levels, aligning with inconsistent consumer demand, and navigating the challenges posed by the limited shelf life of perishable products. The complexities of supply chain logistics add to these difficulties, and to address them effectively, a comprehensive and strategic approach to waste management should be adopted, incorporating both logistical finesse and market adaptability.

Technological innovations and business strategies are important in building resilient food supply chains, which are crucial for reducing food waste. Sharma et al. (2023) emphasize these factors' importance in enhancing supply chains' efficiency and responsiveness, thereby reducing the likelihood of food waste due to supply-demand mismatches or logistical inefficiencies.

Proactive operational planning is also key in combating food waste in retail. This entails tailoring demand forecasts and planning assortment sizes, as noted by Riesenegger and Hübner (2022). The heart of this strategy lies in utilizing comprehensive databases and information systems that are instrumental in minimizing waste by allowing for more accurate and responsive decision-making processes.

Markdown strategies and improved inventory management are practical approaches that many retailers adopt to address the issue of perishables nearing the end of their shelf life. Chen et al. (2016) and Chung (2019) discuss the application of discounts to products close to expiration. This practice, typically involving 20% to 50% discounts, aims to increase the sale of these products before they become unsalable, as elaborated by Chung and Li (2014).

Sales promotion methods, such as "buy one, get one free" (BOGOF), play a multifaceted role in food waste management. Research by Wu and Honhon (2023) indicates that while such promotions might encourage consumer waste, they can benefit retailers. These schemes may decrease overall waste by balancing the waste generated at the consumer level with reductions in retail waste.

Aside from these internal strategies, many retailers engage in collaborative efforts, such as donating unsold but edible food to charities or disposing for animal feeding (Riesenegger and Hübner, 2022). These actions not only help in reducing waste but also align with broader sustainability goals and societal responsibilities. By working with supply chain partners, retailers can address waste reduction at the source, thereby optimizing the entire chain from production to consumption. Nevertheless, Riesenegger and Hübner (2022) establishes that a preferable option should be to prevent leftover food in the first place rather than just trying to manage the waste valorization.

In summary, retailers' strategies to manage food waste are diverse, encompassing everything from operational adjustments and promotional tactics to technological advancements and collaborative initiatives. These strategies are essential to mitigating the negative impacts of food waste and contributing to the sustainability and efficiency of the retail sector and the broader food supply chain.

2.5 Consumer Behavior and Perishables

Consumer purchasing behaviors critically influence the sustainability of food systems, especially regarding perishable goods. Research by scholars such as Horoś and Ruppenthal (2021), Chung (2019) , and Huang et al. (2021) delves into these behaviors, shedding light on their contribution to food waste and the broader implications for sustainability.

Horoś and Ruppenthal (2021) identifies expiration dates, spoilage, consumer over-purchasing, and retailer over-ordering as key factors leading to food waste within grocery retail environments. They advocate for strategies such as sales optimization, food surplus redistribution, and timely price reductions to mitigate waste on nearly expiring products. The study also highlights consumer tendencies to overbuy and preference for products with longer shelf life. This behavior, coupled with the expectation of having a full range of products available even just before the grocery retailer closing time, forces store owners to overstock by around 30%. The study suggests that although store owners believe it is challenging to influence consumer behavior, they expect food industry innovations, like the introduction of indicators of product freshness, to change consumer behaviors.

By examining preferences for citrus fruits, Huang et al. (2021) establishes a consumer emphasis on freshness and appearance rather than size. Their findings suggest that freshness indicators and traceability certification can enhance the willingness to purchase sub-optimal foods, thus reducing waste. The study emphasizes the importance of consumer education and transparent labeling to improve acceptance of such foods since most customers check expiration dates and prefer perishable products with a longer shelf life. This study, along with others (Li et al. (2020b); Harcar and Karakaya (2005); Konuk (2018)), underscores the critical role of freshness cues, such as harvest and packaging dates, in guiding consumer purchases. Moreover, according to Hall-Phillips and Shah (2017), unclear shelf life labels lead to frustration and confusion, emphasizing the importance of clear labeling.

Chung (2019) investigates the effects of varying pricing strategies on consumer demand for perishable foods in the retail sector. His findings suggest that dynamic prices may be more or less effective than a standard "no discount" policy in managing perishable items. Supporting this, Chang and Su (2022) and Le Borgne et al. (2018) have identified that the gradual reduction of prices over a product's shelf-life and the clear communication about low shelf-life discount labels are critical strategies to enhance the appeal of perishables and prevent waste. According to

Le Borgne et al. (2018), explaining to consumers why the low shelf life discount label exists is essential. These discounts aim to make their purchases more appealing and prevent these items from going to waste. Thus, it is unrelated to the product's lack of quality or suitability for consumption.

In summary, consumer behavior plays a pivotal role in the sustainability of perishable goods, and a multifaceted approach is required to address its impact effectively. Key strategies such as refining retail practices, educating consumers, innovating labeling and packaging, and improving price assignment to perishables near expiration date are crucial. These measures directly affect how consumers buy, use, and dispose of food and are critical to reducing food waste and improving food system sustainability. Through these collaborative efforts, it should be possible to significantly change consumer habits to support more sustainable food consumption patterns.

2.6 Advancing Retail Pricing Strategies

Traditionally, retail pricing for perishables has relied on traditional methods like cost-plus, value-based, and inventory-sensitive pricing. Despite their solid foundation, these strategies often need help to adapt to the fast-paced market, especially for perishables, where timely price adjustments are crucial due to their short shelf life.

Machine learning (ML), a subset of artificial intelligence (AI), emerged as a game changer for this. By analyzing vast amounts of data, ML enables the prediction of optimal pricing points, maximization of profits, and enhancement of customer satisfaction, offering a substantial improvement over traditional models. This innovative approach is adept at navigating the complexities of modern markets, including fluctuating consumer behavior, competitive actions, and changes in product quality.

Chen et al. (2018), critiques traditional pricing strategies, advocating for dynamic pricing models empowered by Q-learning mechanisms. These models consider diverse factors such as uncertain demand, customer preferences, and competitive landscapes, allowing for more agile and effective pricing decisions that can significantly boost retailer revenues.

Additionally, integrating blockchain technology into retail systems, as discussed by Zhao and Li (2023), represents another frontier in the evolution of pricing strategies. Blockchain-based traceability systems (BTS) in dual-channel markets for perishables can minimize product losses and provide verifiable product quality information, enabling retailers to justify premium pricing models.

These studies indicate the shift from traditional to more advanced, data-driven pricing methodologies, where machine learning stands out as a key driver of this evolution, enabling the creation of dynamic, adaptable pricing strategies that can consider an extensive range of market factors. Adopting these technologies not only addresses the limitations of previous models but also paves the way for more sophisticated and effective pricing mechanisms.

2.7 Dynamic Pricing for Perishables

Dynamic pricing emerges as an important strategy that enables businesses to adjust prices in real-time based on factors such as demand, supply, market conditions, and customer behavior. Fueled by data analytics and technological advancements, it is increasingly adopted across various sectors to optimize pricing for revenue maximization and market competitiveness. This empowers businesses to swiftly respond to market changes, enhancing their ability to leverage opportunities and mitigate risks associated with price sensitivity and fluctuating demand.

Pioneering research by Chung and Li (2014) underscored the benefits of a multi-period pricing policy over traditional single or two-period policies, particularly regarding profit maximization and food waste reduction. However, their study did not account for menu costs, which Chen et al. (2018a) later identifies as a crucial oversight. The authors found that, while ignoring menu costs leads to a decrease in optimal prices over time for perishable products, incorporating moderate menu costs suggests that a single price adjustment could be more beneficial for retailers.

Innovative approaches to dynamic pricing further include studies by Buisman et al. (2019), who investigated the effects of individual or combined use of dynamic pricing and dynamic shelf life on various retail outcomes, and Chung (2019), who examined the impact of displaying products based on remaining shelf life and applying discounts accordingly. The author underlines the complexity and potential of integrating multiple strategies for managing perishables.

Research by Scholz and Kulko (2021) stresses that dynamic pricing of perishable food based on freshness can increase revenue and reduce food waste, making it a promising strategy for sustainable business models, underscoring the importance of integrating product perishability into pricing strategies, highlighting the potential for dynamic pricing to contribute significantly to sustainability and efficiency in sectors dealing with perishable goods. Sanders (2023) further discusses dynamic pricing and organic waste landfill bans as solutions to reduce grocery-retail food waste based on a structural model with perishability data. The author notes that analyzing these alternative scenarios reveals that dynamic pricing dominates an organic waste ban.

Furthermore, studies on dynamic pricing for perishables have explored various models and their implications for supply chain management and profitability. For instance, a study by Zheng et al. (2021) found that optimal prices for new products are positively related to potential demand and negatively related to the decay rate. In contrast, prices for old products have a positive relationship with both. Ordering decisions are unrelated to the quantity of old products (Zheng et al., 2021). This insight reveals the complexity of pricing perishables, where the interplay between product freshness, demand, and supply chain decisions must be carefully balanced to achieve optimal outcomes.

Despite the promising outcomes, the studies that have made the most significant advancements were by applying reinforcement learning techniques, such as the Q-learning algorithm, due to their flexibility and adaptiveness. Kayikci et al. (2022) and Chen et al. (2018b) demonstrate the application of reinforcement learning in dynamic pricing, highlighting its ability to learn from past data without needing complete knowledge of the environment. Kayikci et al. (2022) proposed a stochastic dynamic programming method for a four-state pricing strategy (freshness, less fresh/cheaper, redistribution, and disposal Stages) of bulk apples. They used real-time sensor data to determine the freshness of food. The algorithm updates the product's unit price at the beginning of each phase while accounting for the remaining inventory of the entire selling period. Chen et al. (2018b) employed a multi-agent model-free Q-learning algorithm to set a dynamic pricing strategy for perishable products, considering the competition, customer preferences, uncertain demand, and perishable characteristics. They assumed three categories of customers (with varying preferences about retailer distance and price sensitivity).

Moreover, Rana and Oliveira (2015) contribute to this domain by applying model-free Q-learning algorithms for managing prices of interdependent perishable products (flights). Yang et al. (2022) introduced a deep reinforcement learning model to determine the price strategy for maximizing the retailer's profit and minimizing food waste. Both studies further illustrate the potential of reinforcement learning in dynamic pricing strategies.

In summary, exploring dynamic pricing for perishables reveals a complex landscape where businesses can significantly benefit from adaptive pricing models. The integration of reinforcement learning techniques represents a promising direction for future research, offering the potential to navigate the challenges of perishability, demand fluctuations, and competitive pressures more effectively.

2.8 Reinforcement Learning and Applications

Reinforcement Learning (RL) is a branch of machine learning (ML) where an agent learns to make decisions through interactions with its environment. The agent selects actions to maximize cumulative rewards over time, balancing exploration (trying new actions) and exploitation (choosing known profitable actions). RL differs from supervised learning as it does not rely on labeled datasets but instead learns from its environment through feedback in the form of rewards or penalties.

A widely used framework for structuring RL problems is the Markov Decision Process (MDP), which defines how an agent interacts with its environment through a cycle of states, actions, and rewards. The agent's objective in an MDP is to select actions that lead to favorable outcomes by maximizing cumulative rewards, subject to constraints in the environment. RL algorithms utilize this framework to approximate optimal policies—decision-making rules that map states to actions—using methods such as value-based and policy-based approaches.

In recent years, RL has been applied across various domains. Drawing from the systematic review conducted by Sivamayil et al. (2023), some applications are highlighted below:

- Finance: RL is applied to optimize trading strategies, portfolio management, and risk assessments by learning from market dynamics to make profitable decisions.

- Robotics: Robots are taught complex motor skills and environment adaptation using RL, essential for tasks like manipulation, autonomous control, and locomotion.
- Healthcare: RL helps optimize intervention sequences in diagnosis and treatment planning, tailoring interventions to individual patient needs and promoting long-term health outcomes.
- Gaming and Simulation: RL enables breakthroughs in AI for video games, allowing agents to outperform humans in complex games such as Go, Dota 2, and chess.
- Energy Systems: RL is used to optimize energy usage in smart grids and home energy management systems, enhancing efficiency and sustainability.

2.8.1 Markov Decision Process (MDP) Framework

The Markov Decision Process (MDP) is a mathematical framework essential to reinforcement learning, allowing the agent to make sequential decisions in a structured environment. As shown in Figure 2.1, the agent interacts with the environment by observing its current state, taking actions, and receiving rewards based on the outcomes. This iterative process forms the foundation of decision-making in MDPs.

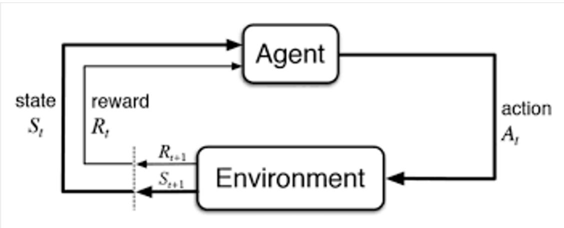

Figure 2.1: Agent-environment interaction cycle in an MDP.

An MDP is typically defined by the following elements:

- **States** (S): The set of all possible states represents every conceivable situation in which the agent may find itself. In retail, a state could represent the current inventory levels of different products, customer demand patterns, or promotional events. Each state provides the agent with the necessary information to make decisions, such as when to reorder stock or adjust prices based on demand.
- **Actions** (A): This is the set of possible actions the agent can take to interact with its environment. In a retail setting, actions might include adjusting product prices, placing new orders for inventory, offering discounts, or launching targeted promotions. Each action affects sales, inventory levels, and customer satisfaction, impacting overall business performance.
- **Rewards** ($R(s,a)$): A reward function assigns a numerical value to each state-action pair, representing the immediate benefit or cost of taking action a in state s. For instance, in retail, the reward could be the immediate profit from sales, customer satisfaction scores, or even

long-term customer retention metrics. The agent's goal in an MDP is to maximize cumulative rewards, which in retail often translates to maximizing revenue, profit, or customer loyalty.

- **Constraints**: Constraints define any additional conditions or limitations that the agent must respect when selecting actions. In retail, constraints could include budget limits, warehouse capacity, supplier lead times, or contractual agreements on minimum or maximum order quantities. These constraints ensure that the agent's actions are practical and align with business rules and resource limitations.

An agent in an MDP uses a policy ($\pi(a|s)$), which maps states to actions to maximize its expected cumulative reward, also called the "return." Through the MDP framework, reinforcement learning models iteratively adjust their policies, aiming for optimal decisions within the constraints of the environment, as depicted in Figure 2.1.

2.8.2 Selected Algorithms and some past Applications in Literature

The next subsection will explore the selected algorithms within RL that will be used in this research, specifically focusing on Trust Region Policy Optimization (TRPO), Proximal Policy Optimization (PPO), Recurrent Proximal Policy Optimization (R-PPO), Advantage Actor-Critic (A2C) and briefly mentioning some of their past applications.

2.8.2.1 Trust Region Policy Optimization (TRPO)

Trust Region Policy Optimization (TRPO) is a reinforcement learning algorithm designed to optimize extensive, nonlinear policies, such as neural networks, focusing on maintaining stability and performance across various tasks. TRPO achieves this by enforcing a trust region constraint during the policy update process, ensuring that each update results in a consistent, step-by-step improvement in policy performance. This monotonic improvement means that each update improves or maintains performance but never decreases it. This approach prevents overly aggressive updates that could destabilize the learning process, making TRPO particularly effective in tasks that require robust, reliable performance without extensive hyperparameter tuning Schulman et al. (2017a).

Building on the solid foundation laid by TRPO, subsequent research has been dedicated to developing algorithms that retain its benefits while simplifying implementation and enhancing sample efficiency. One such algorithm is Proximal Policy Optimization (PPO), which has emerged as a widely adopted approach due to its balance between simplicity and performance. The following section will delve into the specifics of PPO, exploring its applications and the exciting advancements it brings over earlier methods like TRPO.

2.8.2.2 Proximal Policy Optimization (PPO)

Proximal Policy Optimization (PPO) is a reinforcement learning algorithm that has gained considerable attention for its practical balance between simplicity and performance. Initially developed to improve the Trust Region Policy Optimization (TRPO) algorithm, PPO operates by alternating between sampling data from the environment and optimizing a surrogate objective function through stochastic gradient ascent. This approach allows for several minibatch updates, improving sample efficiency compared to traditional policy gradient methods. One of PPO's key innovations is its clipped surrogate objective, which prevents excessive policy updates and helps maintain stable learning (Schulman et al., 2017b).

PPO's effectiveness has been proven in various domains. For instance, it has been used in energy management systems for electric vehicles and photovoltaic (PV) storage units, significantly reducing daily energy costs—by up to 54% when compared to traditional strategies (Alonso et al., 2023). PPO has also proven valuable in autonomous collision avoidance systems in multi-ship environments and optimizing communication systems for 6G networks(Rongcai et al., 2023; Saikia et al., 2023). These examples demonstrate the algorithm's versatility and capability to handle complex tasks in simulated and real-world settings.

2.8.2.3 Recurrent Proximal Policy Optimization (R-PPO)

Recurrent Proximal Policy Optimization (R-PPO) extends the Proximal Policy Optimization (PPO) framework by incorporating recurrent neural networks (RNNs) to handle environments characterized by temporal dependencies and partial observability. This extension is particularly useful in scenarios where the current state depends on a sequence of prior states, making traditional PPO less effective. R-PPO typically integrates Long Short-Term Memory (LSTM) layers to process sequential data, allowing the model to capture and utilize temporal patterns more effectively.

A notable application of this approach is in UAV autonomous landing scenarios, where the RPO-LSTM algorithm (an extension of R-PPO) has demonstrated significantly higher success rates under partial observability than traditional methods. This success underscores R-PPO's robustness and effectiveness in handling complex, dynamic environments with insufficient information (Aikins et al., 2024).

Overall, R-PPO's architecture, integrating LSTM layers with PPO, makes it well-suited for tasks where comprehending and predicting the evolution of states over time is critical, thereby broadening the applicability of PPO in more complex and dynamic environments.

2.8.2.4 Advantage Actor-Critic (A2C)

Advantage Actor-Critic (A2C) is a reinforcement learning algorithm that balances computational efficiency and performance by combining the strengths of both policy-based and value-based methods. It operates within an actor-critic framework, where the "actor" updates the policy by selecting actions, while the "critic" evaluates these actions using the advantage function. The advantage function quantifies how much better or worse an action is compared to the average action at a given state, stabilizing the learning process and improving convergence. A2C is particularly effective in environments where balancing exploration and exploitation is crucial, enabling efficient and stable policy optimization.

A2C has been successfully applied in various fields. For example, it has been used to enhance resource management in network slicing by accounting for user mobility, significantly improving resource allocation in dynamic environments. Additionally, A2C has been applied to energy-efficient beam selection in mmWave networks, optimizing transmission power while maintaining network coverage, thereby boosting Energy Efficiency (EE) in wireless communications (Li et al., 2020a; Dantas et al., 2023).

2.9 Summary

The literature review delves into the intricate dynamics of food waste management, consumer behavior, and dynamic pricing strategies within the retail sector, focusing on perishables. A significant opportunity exists for innovation in applying machine learning (ML) techniques to improve dynamic pricing models, directly aligning with the objectives of this research. Unlike traditional dynamic pricing models, which adjust prices in real time based on demand, supply, and shelf-life, this research seeks to refine the frequency and regularity of price updates to identify effective discount ranges that encourage consumer purchases of perishable products.

The literature identifies dynamic pricing as a potent tool for reducing food waste and increasing retailer profits. However, it also highlights a critical gap in applying ML techniques, particularly in the context of perishables, where the characteristics of the products and the consumer response to price changes over time are not adequately addressed. Furthermore, it underscores the necessity of integrating consumer behavior insights into dynamic pricing models. This gap presents an opportunity for this research to contribute to the field. The study aims to employ reinforcement learning techniques to price perishable products dynamically. This approach will consider key factors such as expiration dates, consumer purchasing behavior, and demand fluctuations to optimize the timing frequency and the magnitude of price adjustments. The goal is simultaneously optimizing waste reduction and profit maximization while determining the most effective discount range for encouraging consumer purchases.

The literature review solidifies the investigation into ML-based dynamic pricing strategies for perishables by aligning these insights with the research objectives. The methodology will focus on developing and testing a model that considers the need for more frequent price updates and seeks to identify the optimal intervals for adjusting discounts or prices.

Conclusively, the literature review reinforces the importance of dynamic pricing in tackling perishable product management challenges and delineates a potential contribution to this research. By exploring the potential of reinforcement learning techniques to refine the regularity and effectiveness of pricing strategies, the research aspires to deliver innovative solutions that meet the modern retailer's needs, fostering sustainability, reducing food waste, and, most importantly, improving economic outcomes.

Chapter 3

Methodology and Implementation

This chapter explores the methodology and implementation strategies used to develop and evaluate reinforcement learning (RL) models to optimize pricing strategies to reduce food waste and increase profitability.

The research starts by preparing the environment and preprocessing collected data. The datasets include product information, stock levels, product labels, stock receipts, and comprehensive sales records from the retailer involved in the BeFresh project.

Following data preparation, several RL models — Proximal Policy Optimization (PPO), Trust Region Policy Optimization (TRPO), Advantage Actor-Critic (A2C), and Recurrent PPO (R-PPO)— were chosen to be implemented for their effectiveness in sequential decision-making for retail pricing.

Once the data was prepared, the models were trained in a simulated retail environment using the OpenAI Gym framework, a testament to their adaptability. This framework supports flexible and scalable simulations, allowing the models to learn and adjust pricing strategies dynamically based on current sales data and product shelf life. This adaptability is a crucial feature for the models, ensuring they can handle the ever-changing landscape of retail pricing.

Extensive experiments, including hyperparameter tuning and reward calculation, are conducted to ensure the models' robustness. The experimental setup and testing procedures validate the proposed pricing strategies. Additionally, detailed information about the computational resources and configurations is provided.

This overview covers the main components and processes of the methodology and implementation, laying the groundwork for the detailed descriptions in the following sections.

3.1 Data Collection and Preparation

This section will outline the methodologies employed in collecting, cleaning, and preparing the data for analysis. This encompasses identifying the data sources, the procedures undertaken to handle and process the data, and the criteria for selecting relevant features for the chosen models.

3.1.1 Datasets

The retailer collaborating with the BeFresh project provided five data sources, each contributing essential information for the research. The datasets are related to perishable products, focusing on the yogurt portfolio. They contain detailed information on product attributes, stock levels, product labels, stock receipts, and comprehensive sales records for various yogurt products. Each dataset is described in detail in the sections that follow.

3.1.1.1 Labels

This dataset monitors product pricing and sales across various stores, detailing discounts, sales status, and relevant dates. It comprises 1,966,781 records collected from February 28, 2021, to March 17, 2022. Of these records, 997,239 correspond to unsold perishable products. The dataset includes information from 383 stores and covers 1,086 SKUs.

3.1.1.2 Sales

The sales dataset encompasses comprehensive product information, sales quantities (both regular and promotional), stock status, and financial metrics, including prices and purchasing costs. This dataset spans from December 1, 2021, to March 14, 2022, and comprises a total of 8,802,140 records. Each record represents a specific product on a particular date across various stores. From these records, 5,904,292 records indicate products that were not sold. This extensive dataset provides a robust foundation for analyzing sales patterns, stock management, and financial performance within the given timeframe.

3.1.1.3 Stock

This dataset provides detailed tracking of inventory levels for various perishable products, identified by SKU, across different stores. Each record in this dataset represents the stock on hand for a specific product at a particular store, along with the date and time of the recorded information. Spanning from December 1, 2021, to May 30, 2022, this dataset comprises a total of 15,486,942 rows.

3.1.1.4 Stock Received

The stock received dataset tracks inventory information for various perishable products, identified by SKU, across different stores. Each row in the dataset represents the quantity of a specific product received at a particular store, along with relevant dates, including the validity date and the date of receipt. Spanning from February 27, 2020, to March 20, 2022, the dataset comprises a total of 3,781,500 rows.

3.1.1.5 Articles

The articles dataset provides detailed product information, including SKU numbers, descriptions, classifications, department and subclass codes, shelf life, brand information, capacity units, and weight. With a total of 10,094 rows, this dataset is instrumental for inventory management, product categorization, and detailed tracking of product attributes.

3.1.2 Data Preprocessing

The data cleaning and preparation involved several steps to ensure data integrity and usability for subsequent analysis. All datasets were filtered to include records between December 1, 2021, and March 14, 2022, as this timeframe was standard across all datasets. This selection ensured consistency and alignment in the data, reducing potential discrepancies. Additionally, null values in the datasets were checked and handled accordingly.

In the *Labels* dataset, a new column was introduced to calculate the validity period, defined as the time between placing a discount label and removing the product from the shelf. Two days were added to the calculated validity period to account for early product removal due to waste (i.e., when a product was not purchased and removed by the store). Records with a negative validity period were excluded from the analysis.

The dataset was cleaned to ensure the labeling date consistently preceded the sale date. In cases where the labeling date occurred after the sale date, the labeling date was replaced by the sale date. The remaining shelf life at the time of sale was also calculated and adjusted by adding two days. Observations where the expiration date was earlier than the labeling date were excluded from the analysis. To capture the entire period during which a product was available, additional rows were generated for each day between the labeling date and the product's removal from the shelf. This approach resulted in a more granular dataset with daily records.

After proper sorting, the *Labels* dataset was merged with the *Sales* dataset. Missing values in the 'price_no_label' column were inputted with the corresponding 'oldpvp' values, while other missing values were addressed by filling them with the mean price for each combination of SKU and store. Only combinations with more than thirty sales were retained to ensure sufficient data for analysis.

The *Stock* and *Stock Received* datasets were loaded, filtered, and merged with the previously resulting dataset to create a comprehensive dataset for the models. Missing values in the datasets were handled by imputing them with appropriate defaults.

Corrections were made for identified errors, particularly erroneous date validation values. The remaining shelf life for each product was calculated, incorporating stock receptions and sales without labels.

Finally, the cleaned and prepared dataset was saved for further analysis. Additional aggregation and visualization were performed to explore key metrics, including mean remaining shelf life and total units sold, grouped by product subcategories. The processed data was then saved for subsequent analysis and modeling. The final dataset consisted of 70,250 rows following feature selection, providing a robust foundation for predictive modeling and analysis.

The model's feature selection was guided by predicting product sales (more detail in section 3.2.2.3) and understanding the factors influencing consumer purchasing behavior that are influenced by the discounts applied to the items. The selected features are:

- **id:** Unique identifier for each record in the dataset.
- **sku (Stock-Keeping Unit):** Uniquely identifies each product.
  Enabling precise tracking of individual product performance and sales patterns.
- **idstore:** Unique identifier of each store.
  Capturing variations in sales performance due to location, customer demographics, and store characteristics.
- **date:** The date of the record in the format YYYY-MM-DD.
  Essential for time series analysis, identifying seasonal trends, and understanding sales fluctuations over time.
- **price_no_label:** The price sale of the product (without any discount).
  Essential for calculating the "new price" after a discount is applied or for estimating waste costs.
- **discount_percentage:** The assigned discount percentage from the store.
  Important for understanding the pricing strategy and its influence on consumer demand.
- **discounted_price:** The price value after discount for the product.
  Showing the price after applying a discount is crucial for evaluating the effectiveness of promotional strategies.
- **days_to_expiration_no_label:** The product price without a label.
  Represents the regular price of the product, serving as a baseline for comparing the impact of discounts on sales.
- **days_to_expiration_with_label:** Indicates the remaining shelf life of the product with a label, measured in days, when a product is sold, influencing consumer purchasing decisions.
- **sales_regular:** The total number of units sold without a discount label.
  Serves as a baseline support for the sales performance of a product without promotional influence.
- **c_label_sold:** The total number of units sold with a discount label. Helps to quantify the impact of sales on products with discounts.
- **c_sold_no_label:** The total number of units sold without a label.
  Providing a comparison for understanding the impact of discounts versus regular sales.

- **inventory_level:** The available product stock (without discount) at the store level. Inventory level is key for understanding stock availability and its relation to sales performance and promotional effectiveness.

- **sales_velocity:** The rate at which a product is sold over a given period. Provides insights about the speed at which inventory is turned over and how promotions may accelerate the process.

- **profit_margin_percentage:** Percentage of profit achieved on a product sale. Helps to evaluate the financial impact of discount strategies.

- **prod_cat:** The product category to which the item belongs.

- **prod_subcat:** The subcategory within the main product category.

- **aux:** Counts the number of consecutive rows that share the same ID.

The selected features were chosen for their relevance in capturing key aspects of the sales process, including product identification, pricing, promotional activities, inventory status, and temporal factors. This comprehensive feature set enables the model can effectively analyze and predict sales patterns, understand promotional impacts, manage inventory, and support strategic decision-making in sales and marketing, and therefore address the research question

3.2 Model Development and Implementation

This section outlines the development and implementation of the reinforcement learning models used to optimize discount policies in food retailing. The models are designed to interact with a simulated retail environment, learning optimal pricing strategies in order to maximize profits and minimize food waste.

3.2.1 Environment Setup

Setting up the environment is a key part of developing and evaluating reinforcement learning models for optimizing discount policies in food retailing. This subsection discusses the tools and libraries used, the steps taken to prepare the data, and the creation of the simulation environment.

3.2.1.1 Tools and Libraries

The computational environment for this research is configured using a combination of specialized libraries and tools designed to support reinforcement learning and data preprocessing tasks. The primary programming language used was Python, due to its extensive ecosystem of libraries and frameworks suitable for machine learning and data analysis. The following key libraries and tools were utilized included:

- **NumPy and Pandas**: These libraries were chosen for their robust capabilities in data manipulation and processing, allowing efficient handling of large datasets and facilitating the necessary data transformations.
- **Matplotlib and Seaborn**: Used for data visualization, enabling the creation of detailed plots and charts that help in understanding data distributions and patterns.
- **Scikit-learn**: This library provided additional machine learning algorithms and data preprocessing techniques, such as the OneHotEncoder for categorical data, enhancing the preprocessing pipeline and supporting model development.
- **TensorFlow and PyTorch**: These deep learning frameworks were used for building and training the neural network models. Their integration with Stable-Baselines3 allowed for the implementation of advanced reinforcement learning algorithms.
- **Stable-Baselines3**: This library was used to implement the reinforcement learning algorithms.
- **OpenAI Gym**: OpenAI Gym was employed for providing standard reinforcement learning interfaces and environments, ensuring consistency and standardization in the environment setup.
- **Gymnasium (gym)**: This was used for creating the custom environment specific to the retail pricing simulation, allowing the reinforcement learning agents to interact with a simulated retail setting.

3.2.1.2 Data Preparation

The dataset used in this study includes detailed records of sales, pricing, inventory levels, and product shelf life from various retail stores. This data was sourced from multiple datasets, carefully combined, and merged into a final comprehensive dataset, as described in section 3.1.2.

The following key data preparation steps were done:

- **Loading the Dataset:** The final dataset was loaded from a CSV file into a Pandas DataFrame to facilitate easy data manipulation and analysis and sorted to maintain chronological order, which is essential for accurately modeling time-dependent processes.
- **Datetime Conversion:** To ensure consistency in time-based operations, the 'date' column was converted into a DateTime format. This is essential for accurate time-based operations such as sorting, filtering, and modeling processes that rely on temporal data.
- **One-Hot Encoding:** For categorical variables such as 'SKU' (stock-keeping units), a OneHotEncoder was used to transform these categorical values into a format that machine learning models can effectively utilize. This transformation ensures that categorical data is appropriately represented without introducing bias from ordinal relationships.

- **Train-Test Split:** The dataset was split into two sets: 70% for training and 30% for testing (after pre-processing). This division allows the model to learn from one portion of the data and then be validated on unseen data to ensure the generalizability of the results.

3.2.2 Markov Decision Process (MDP) Formulation

This section details the formulation of the Markov Decision Process (MDP) used to model the reinforcement learning environment. States, actions, rewards, and constraints characterize an MDP. In this environment, which simulates a retail scenario with perishable products, the MDP framework allows the agent to make optimal decisions regarding discounting, selling, and restocking. The agent's primary objective is to maximize long-term profit while minimizing waste, achieved through sequential decision-making over time.

3.2.2.1 States

The state space defines all the information available to the agent at each step required to make informed decisions, encapsulating the environment's characteristics that influence decision-making. The state S_t at time step t is represented by a vector of features that describe the current status of the product and the environment. The state vector features include inventory levels, current and original prices, time until expiry, historical sales, the store identifier, and SKU features. The SKU features are one-hot encoded using the OneHotEncoder, providing a unique representation for each product.

3.2.2.2 Actions

The action space aims to allow the agent to dynamically adjust prices for products with short shelf life, manage inventory, and ensure product availability through restocking. This enables it to learn optimal policies for profit maximization and waste minimization.

At each time step t, the agent takes an action A_t from the action space to influence the environment. The action space in this environment is multi-discrete, allowing the agent to choose from a set of possible actions across three dimensions:

1. **Discount Level**: The agent can decide the discount level from a predefined set of values: $\{30\%, 40\%, 50\%, 60\%, 70\%\}$. These values represent the percentage discount applied to the product's original price. The chosen discount level directly affects the product's demand.

2. **Units to Sell**: The agent determines the number of units it aims to sell at the discounted price. For discounted products, the available inventory and the product's demand under the chosen discount influence the number of units sold. For non-discounted products, the bigger the remaining shelf life, the bigger the demand, as long as the inventory available allows it.

3. **Restocking Decision**: The agent makes a binary decision regarding whether he should restock or not:

- 1 for restocking the product.
- 0 for not restocking.

If restocking is selected and the product's inventory level is zero, the inventory is replenished according to the historical median inventory for the SKU at that store ID.

3.2.2.3 Rewards

The reward function $R(S_t, A_t)$ aims to guide the agent's learning by assigning a numerical value to each state-action pair. The reward is designed to balance two goals: increasing sales of perishables through discounts and minimizing waste. It is composed of two main components:

- **Sales**: The revenue generated from selling units. The profit depends on the number of units sold and the price (original or discounted) at which they are sold. For each unit sold, sales are calculated as:

$$\text{Sales} = \text{Price} \times \text{Units Sold} \tag{3.1}$$

  Higher discounts increase sales volume but decrease profit per unit, creating a balance between sales and revenue.

- **Waste Cost**: Waste cost is determined by the amount of unsold inventory. Unsold units nearing expiration (two days) generate waste. The waste cost is incurred for any units left unsold, particularly when their shelf life is close to expiration. Waste is calculated as:

$$\text{Waste Cost} = (\text{Original Price} \times 0.30) \times \text{Unsold Units} \tag{3.2}$$

  The calculation with the factor of 0.30 represents the acquisition cost per unit of waste. With this penalty, it's expected that the agent feels discouraged from allowing products to expire without being sold.

The reward is calculated as follows:

$$R(S_t, A_t) \ = 0.65 \times \text{Sales} - 0.35 \times \text{Waste Cost} \tag{3.3}$$

This reward structure aims to encourage the agent to maximize long-term profitability by selling as many units as possible at an optimal price while minimizing the cost associated with waste.

3.2.2.4 Constraints

The environment incorporates several constraints that the agent must operate within. These constraints aim to enhance the realism of the simulation by restricting the agent's actions to those that align with the practical limitations and operational challenges inherent in managing perishable goods. The constraints are summarized as follows:

- **Inventory Limits:** The agent is not allowed to sell more units than are available in the inventory. There are separate inventory pools for non-discounted and discounted products, and the available stock in each category constrains sales. This reflects the real-world scenario where stock levels limit sales.

- **Demand Estimation:** The environment estimates demand separately for non-discounted and discounted products. Non-discounted products experience higher demand when their remaining shelf life is longer, reflecting consumer preferences for fresher goods. As the number of days until expiration decreases, the demand for these products gradually declines as customers perceive them as less valuable or riskier to purchase. This relationship between demand and product freshness is captured through an expiration multiplier, which scales demand based on the remaining days to expiration.

  For discounted products, demand increases with the application of a discount multiplier, aiming to simulate consumer responsiveness to price reductions. However, the agent's ability to sell excess inventory is limited by these realistic demand patterns. Even if a product is available in large quantities, higher inventory levels do not guarantee higher sales if demand remains low, especially for near-expiry goods. Therefore, the agent must carefully adjust its pricing and discount strategy to align with actual consumer demand, optimizing sales without relying on aggressive discounting or overstocking.

- **Expiration and Transfer to Discount:** Products nearing expiration within five days are automatically moved to the discounted inventory. This simulates standard retail practices where goods nearing their expiration date are discounted to increase the likelihood of selling them before they become waste.

- **Discounted Not Sold Products:** Discounted not sold products are moved into the next day inventory. This aims to simulate retail practices where goods nearing their expiration date are discounted until they become unsaleable.

- **Restocking Limits:** Restocking is constrained by both the inventory level and historical sales data. Restocking is only possible when the inventory level drops to zero. When a restocking decision is made, the quantity restocked is determined based on the historical median of past inventory level for the specific SKU at that particular store. This constraint ensures that restocking aligns with past real-world demand patterns, avoiding overstocking or understocking situations.

- **Termination Conditions:** The episode should terminate if the agent reaches the end of the dataset or if the cumulative reward falls below a predefined threshold (which represents significant financial loss). This constraint aims to prevent the agent from continuing in a scenario where it consistently incurs negative rewards.

3.2.3 Algorithms and Model Architectures

This subsection presents the reinforcement learning (RL) algorithms utilized through the research and the motivation behind their selection. The algorithms discussed include Trust Region Policy Optimization (TRPO), Proximal Policy Optimization (PPO), Recurrent Proximal Policy Optimization (R-PPO), and Advantage Actor-Critic (A2C). These algorithms were carefully chosen based on their strengths, such as handling high-dimensional continuous action spaces, managing partial observability, and ensuring stable learning in complex environments. These algorithms were implemented using the Stable-Baselines3 library, renowned for its reliability and widespread use in reinforcement learning research. In particular, some implementations used the SB3-Contrib module, which offers advanced features tailored to more specialized needs. Leveraging the Stable-Baselines3 toolkit ensured that the computational experiments were based on robust and efficient implementations, further enhancing the reproducibility of the results.

3.2.3.1 Trust Region Policy Optimization (TRPO)

Trust Region Policy Optimization (TRPO) is a policy optimization algorithm in reinforcement learning (RL) designed to maintain stable and reliable learning through constrained policy updates. Unlike simpler policy gradient methods, TRPO introduces a safeguard by limiting the divergence between consecutive policies, preventing drastic changes that could destabilize learning. This constraint is based on the Kullback-Leibler (KL) divergence, which ensures that the learned policy does not change too drastically between updates, allowing for gradual improvements.

In this implementation, TRPO model was configured with a discount factor of $\gamma = 0.99$, determining how much future rewards are valued. A step size of 2048 represents the number of interactions the agent has with the environment before updating its policy. Additionally, the entropy coefficient is set to 0.0, meaning that the agent did not prioritize exploration and instead concentrated on refining its current policy, the learning rate was set to 0.0003. The value function coefficient was set to 0.5, which balances the importance of minimizing the error in the value function versus improving the policy. These parameters were chosen based on common practices in RL and specific characteristics of the TRPO algorithm for training. They balance exploration, stability, and efficiency and are typical for RL environments with continuous control tasks (e.g., OpenAI Gym Environments: CartPole, MountainCar, Pendulum and Mujoco Tasks: HalfCheetah, Walker).

3.2.3.2 Proximal Policy Optimization (PPO)

Proximal Policy Optimization (PPO) is a reinforcement learning (RL) algorithm that balances exploration and exploitation while ensuring stable and efficient policy updates. Unlike traditional policy gradient methods, PPO avoids extremely bold updates by limiting the change in the policy at each step, similar to Trust Region Policy Optimization (TRPO), but with a simpler and more practical implementation. PPO achieves this using a clipped surrogate objective function, which ensures that updates do not deviate excessively from the current policy. This effectively stabilizes the learning and prevents performance degradation.

The configuration used in this implementation includes a discount factor of $\gamma = 0.99$, which determines how much future rewards are weighted. A step size of 2048 specifies how many interactions the agent collects from the environment before updating its policy. The learning rate was set to $\alpha = 0.0003$, a common value that ensures the agent learns efficiently without making updates too aggressively. Other key hyperparameters include a batch size of 256, which controls the number of samples used per training iteration, and a clip range of 0.2, which defines the range within which policy updates are allowed. This clip range is crucial to PPO's design, as it restricts the extent of policy changes to prevent updates that are too large, thereby maintaining stable learning. Additionally, PPO performs multiple updates using the same batch of data (in this case, over 10 epochs), a technique that allows for more efficient use of each interaction with the environment.

The value function coefficient, set at 0.5, controls the trade-off between improving the policy and minimizing the error in the value function. This ensures that the agent balances learning the expected future rewards (value function) with improving the policy itself. The entropy coefficient is set to 0.0, meaning that the agent does not emphasize additional exploration, focusing instead on optimizing the policy based on current knowledge. These parameter choices reflect commonly used settings for PPO in continuous control tasks, particularly from environments like Mujoco and OpenAI Gym, and were chosen after some finetunning.

3.2.3.3 Recurrent Proximal Policy Optimization (R-PPO)

The Recurrent Proximal Policy Optimization (R-PPO) algorithm is an extension of the Proximal Policy Optimization (PPO) algorithm, designed specifically to handle environments with partial observability by incorporating recurrent neural networks (RNNs), such as LSTMs (Long Short-Term Memory networks). In this implementation, R-PPO allows the agent to learn and maintain a memory of past states, which helps it make better decisions in environments where not all information about the current state is observable at each time step.

Several key parameters control the R-PPO model. Like in PPO, the discount factor is set to $\gamma = 0.99$, and the learning rate, $\alpha = 0.0003$, entropy coefficient is set to 0.0, and the value function coefficient is set at 0.5. In recurrent algorithms like R-PPO, the number of steps collected before a policy update is reduced compared to standard PPO. Here, the agent collects 128 steps of experience before each update. This smaller step size helps manage the sequential nature of the data, allowing the recurrent architecture to capture dependencies over time. The clip range, set to 0.2, limits how much the policy can change during each update, which is a critical aspect of R-PPO's stability. In this implementation, the LSTM-based policy allows the agent to maintain memory over time, which is critical for environments where full observability is not guaranteed. The policy's recurrent nature allows the agent to remember past states and actions, improving decision-making in dynamic environments. Similarly, this parameter were chosen based on the specific needs of handling partially observable environments with temporal dependencies, and sources like Stable-Baselines3 and OpenAI's Spinning Up were used.

3.2.3.4 Advantage Actor-Critic (A2C)

Advantage Actor-Critic (A2C) is a synchronous version of the Actor-Critic algorithm that leverages parallel environments to improve training efficiency. The agent learns from both the policy (actor) and the value function (critic) to balance immediate rewards and long-term rewards. Making A2C well-suited for tasks that require balancing multiple objectives, such as dynamic pricing and inventory management, as it efficiently learns to optimize the agent's policy while minimizing errors in value estimation.

Several key parameters are used in this implementation to optimize the performance of A2C. Similar to previous models, discount factor was set to ($\gamma = 0.99$). A learning rate of $\alpha = 0.0007$ is used, which is slightly higher than the default value in PPO but still ensures stable learning. This allows the agent to update its policy efficiently without overshooting the optimal solution.

In A2C, the number of steps (5) determines how many interactions the agent collects from the environment before updating the policy. Smaller steps, like the 5 steps used here, enable the agent to update its policy more frequently, which is beneficial for environments where rapid feedback and updates are necessary. This choice is common in A2C to ensure faster convergence in environments with relatively stable dynamics.

The value function coefficient of 0.5 balances the trade-off between improving the policy and minimizing the error in the value function. These parameter choices reflect sources like the Stable-Baselines3 documentation and OpenAI's Spinning Up guide and commonly used settings, particularly from environments like OpenAI Gym, and were chosen after some fine-tuning.

3.2.4 Training and Test Strategy

The training strategy involved several essential steps, including preparing the training and testing environments. The dataset is divided into training (70%) and testing (30%) sets, ensuring that the model is evaluated on unseen data. For the training dataset, the period spans from the earliest available date (December 01, 2021) up to 70% of the dataset timeline. The testing dataset covers a distinct period from February 17, 2022, to March 14, 2022, with a total of 21,075 rows. This division ensures that the model is evaluated on data it has not encountered during training, thus improving its robustness and generalization ability.

The reinforcement learning models such as TRPO, PPO, R-PPO, and A2C were initialized with specific parameters, including learning rate, discount factor, number of steps, number of training steps, and batch size. These parameters are crucial for model convergence and performance. The models interacted with the environment for 1 million training steps, collecting experiences to update decision-making policies and value assessments. During this process, evaluation steps were conducted every 10 thousand training steps to monitor the model's learning progress and assess its generalization capacity.

The training process is continuously monitored using evaluation callbacks, which periodically assess the agent's performance in a test environment. This evaluation is essential for ensuring that the learned policy generalizes beyond the training data and is robust when faced with unseen scenarios. Additionally, checkpoints are used to save the model's progress at regular intervals, acting as a safeguard against interruptions and providing the option for early stopping if performance stagnates or begins to degrade.

Results from the training and evaluation phases were systematically logged to ensure a comprehensive understanding of the model's progression. Key performance metrics, including average reward, profit, and waste cost, were saved at each evaluation step. These metrics were essential for analyzing the model's ability to optimize pricing strategies for perishable goods. This approach also enabled detailed post-training analysis to determine the most effective hyperparameter configurations.

Following training, the model's performance was evaluated in the test environment over five thousand steps. The reward function, detailed in section 3.2.2.3, aims to play a significant role in the training strategy, balancing profit and waste costs to encourage strategies that maximize profit while minimizing waste. This approach was crucial for the model to learn effective pricing strategies.

3.2.5 Benchmark Setup and Evaluation

A benchmark was designed to evaluate the performance of a fixed discount policy. This was achieved by implementing it using the Proximal Policy Optimization (PPO) algorithm and setting the discount value to a fixed value. The primary goal of this benchmark is to assess the model's ability to optimize discount strategies for perishable products in the retail sector, focusing specifically on a 30% discount.

The PPO algorithm was selected for the benchmark for its stability, performance, and efficiency. Its clipped objective function ensures stable and reliable learning, which seems essential when optimizing discount strategies in a retail environment. PPO is widely recognized for its robustness in reinforcement learning tasks, particularly when decision-making involves balancing multiple, often competing, objectives. The algorithm achieves this stability through its unique policy update mechanism, which prevents large, destabilizing changes by limiting the magnitude of updates. This gradual and controlled learning process is crucial for discount optimization, as sudden shifts in pricing strategies could result in negative outcomes, such as overstocking or excessive discounting.

Another key reason for choosing PPO is its ability to effectively handle discrete and continuous action spaces, making it ideal for simultaneously managing discount levels and inventory decisions. Additionally, PPO is sample-efficient, meaning it can learn effectively from limited data. This is an important factor in perishable product retailing, where data availability is often constrained, and collecting large datasets can be costly and time-consuming. Moreover, PPO strikes a balance between exploration and exploitation. In retail scenarios, it is critical to explore various discount strategies to identify the most effective ones while capitalizing on successful policies already in place. Finally, PPO's widespread recognition in research and industry further reinforces its suitability for this application.

The benchmark setup, which includes configuring the environment, model, and performance tracking mechanisms to ensure a consistent evaluation of the fixed discount policy, is outlined in the section below.

3.2.5.1 Configuration

The benchmark was configured to train and evaluate a fixed 30% discount policy using the PPO algorithm. Although the benchmark, the discount level was fixed to a specific level the agent was required to optimize other decision variables while maintaining a consistent discount strategy. The decision to fix the discount level at 30% was made to establish a baseline for profitability and waste reduction while allowing the model to focus on adjusting inventory, sales, and pricing strategies based on this fixed constraint.

The training and evaluation processes for the PPO model with a fixed 30% discount were conducted in the same manner as the procedures outlined for the other models. The configuration for this implementation adhered to the same parameters as previously specified for the PPO model.

Chapter 4

Results and Discussion

The analysis of total sales (figure 4.1) and profit (figure 4.3) by algorithm reveals significant differences in performance among the evaluated algorithms, TRPO, PPO, R-PPO, and A2C when compared to the benchmark with a fixed discount of 30% discount. By summing the metrics for each algorithm over the evaluation period, the analysis aims to assess the long-term cumulative impact of the algorithms on key retail performance metrics. This approach is particularly suited to the retail industry, where inventory management, pricing, and replenishment decisions are based on cumulative performance rather than short-term fluctuations.

Through the analysis, it is possible to observe which algorithm delivers the highest overall total volume of sales while minimizing waste over time, providing insights into the most effective strategies for optimizing retail operations. As previously mentioned, the results revealed significant differences in performance between the algorithms, with TRPO and benchmark delivering a higher cumulative volume of sales with varying profits. However, this comes with the highest waste in terms of cost (figure 4.6), percentage (figure 4.2), and total units not sold (figure 4.5) suggesting that they may prioritize higher sales at a higher cost, which may result in a larger volume of units sold as it is backed up by figure 4.4, in particular for TRPO. While this contributes more significantly to overall revenue, it also significantly increases waste.

A2C consistently minimizes waste in both costs and units (figures 4.6 and 4.5, respectively) while achieving a decent profit (figure 4.3). This makes it attractive for balancing waste reduction and profitability. As for PPO, it has the lowest profit (figure 4.3) performance but still pushes units, figure 4.4, by offering more discounts, particularly at higher discount categories (figure 4.1).

In conclusion, comparing algorithms reveals distinct trade-offs between profit maximization and waste reduction. TRPO emerges as the most profitable algorithm but at the cost of higher waste, both in terms of total units and associated waste costs. On the other hand, although PPO has increased unit sales due to the higher discounts and minimization waste, it has also resulted in lower overall profits due to extreme discounting.

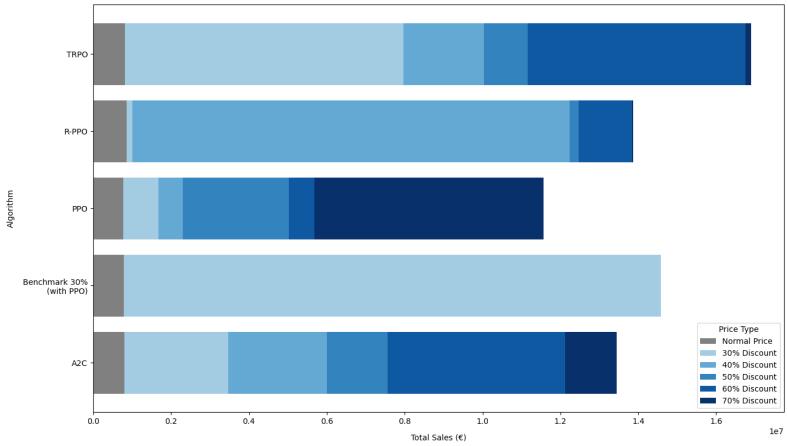

Figure 4.1: Total Volume of Sales in Euros

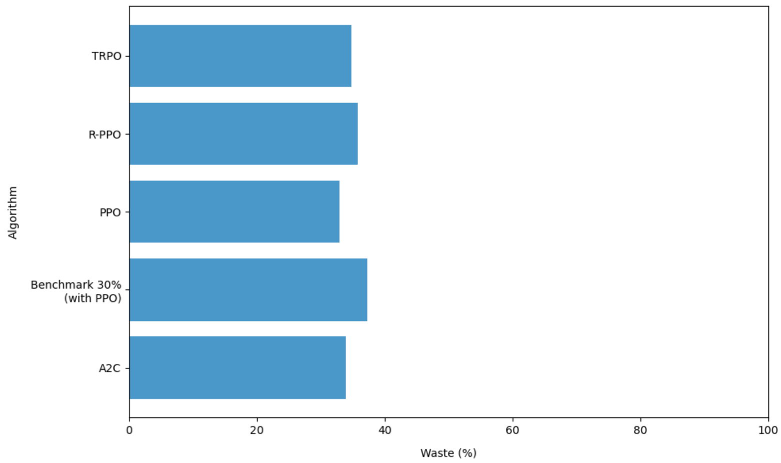

Figure 4.2: Percentage of Waste

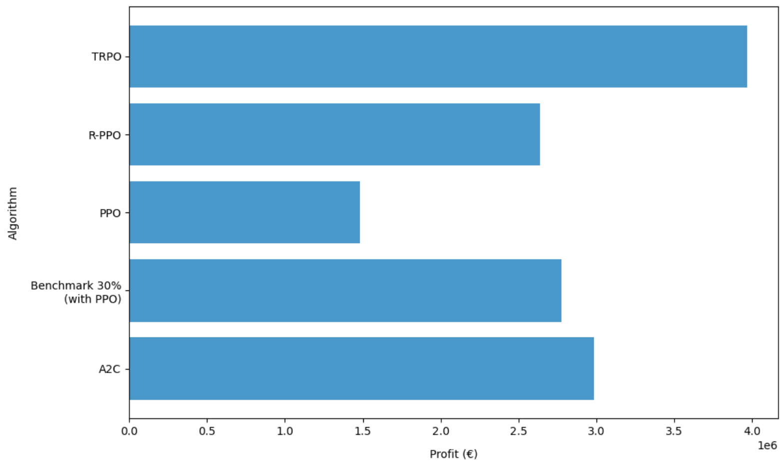

Figure 4.3: Total Profit

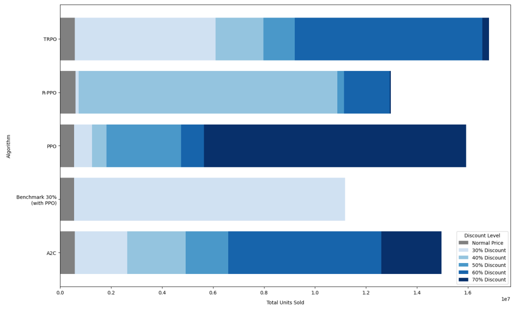

Figure 4.4: Total Units Sold

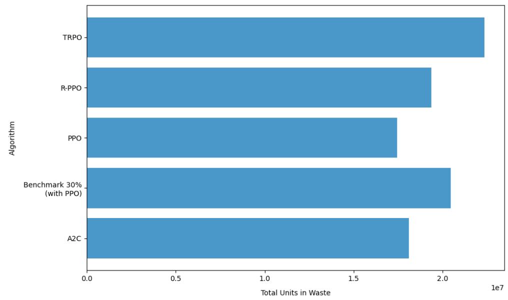

Figure 4.5: Total Units Wasted

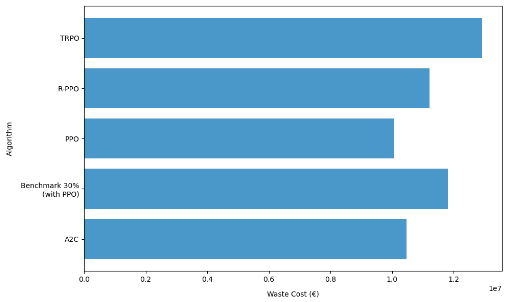

Figure 4.6: Total Waste in Euros

Regarding the research question that originated this investigation, "What is the optimal discount range that encourages consumers to purchase perishables products and maximize retailer profitability?" the present results provide insights into the relationship between discount levels and consumer behavior when purchasing perishable products. Each algorithm applied various discounting strategies, and their outcomes on sales, waste, and profit have revealed significant findings about the optimal discount range for perishable products.

From these findings, it is possible to conclude that the optimal discount range that encourages consumers to purchase perishable products while maintaining a balance between sales volume, profitability, and, in some considerable waste reduction, seems to be between 30% and 60%. This range appears to prove effective for the following reasons:

- Discounts within the 30% to 60% range are sufficient to attract consumers without extreme price reductions. The sales data from algorithms such as A2C, TRPO and the PPO-based Benchmark (with a 30% discount) demonstrate that moderate discounts can still generate substantial sales volumes (figure 4.1) while avoiding the pitfalls of over-discounting when comparing, for instance, the number of units sold versus the total sales (figures 4.4 and 4.1, respectively).

- A higher discount of 70% can indeed boost short-term sales, but it leads to reduced total revenue (figure 4.1), as demonstrated by the PPO algorithm. While this approach successfully moves large quantities of products (figure 4.4), it results in the lowest profit (figure 4.3) due to the steep discounts applied. This contrasts with the discount ranges between 30% to 60%, which offer more balanced results, yielding solid sales figures (figure 4.1) while (in most cases) significantly reducing the number of units wasted (figure 4.5), making them more efficient overall.

In recap, the 30%-60% discount range emerges as the most effective and sustainable option for encouraging consumer purchases of perishable goods and supporting profitability. Although a deeper discount (70%) may offer short-term gains in sales volume, it tends to lead to reduced profitability, making it a less sustainable long-term strategy. On the other hand, the 30%-60% discount range tends to offer a more balanced approach that aligns consumer behavior with retailer goals, ensuring both profitability and waste reduction.

Regarding the performance of different reinforcement learning algorithms, the average reward per step was evaluated, as shown in figures 4.7 and 4.8. Across all algorithms, the rewards exhibit fluctuating patterns, with noticeable spikes at different moments. This suggests that while the agents learn and apply successful strategies at some moments, maintaining consistent performance remains a challenge. None of the algorithms show signs of stabilization, indicating that the environment or task complexity may influence the agents' ability to converge on a stable policy. Additionally, the lack of stabilization can be due to insufficient training time, suboptimal hyperparameters, or a reward function that does not effectively guide the agents toward a consistent strategy. Despite these fluctuations, algorithms like PPO and R-PPO demonstrate comparatively better performance than the remaining, achieving higher rewards and showing more robust learning patterns, though they, too, exhibit variability. Further tuning and adjustments may be required to achieve more stable and consistent results in this environment.

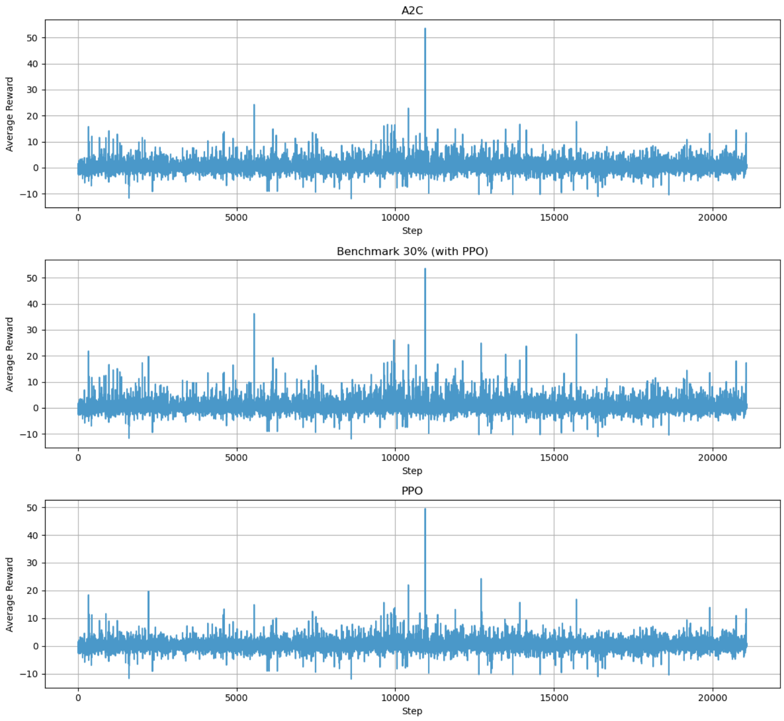

Figure 4.7: Average Rewards for: A2C, Benchmark and PPO

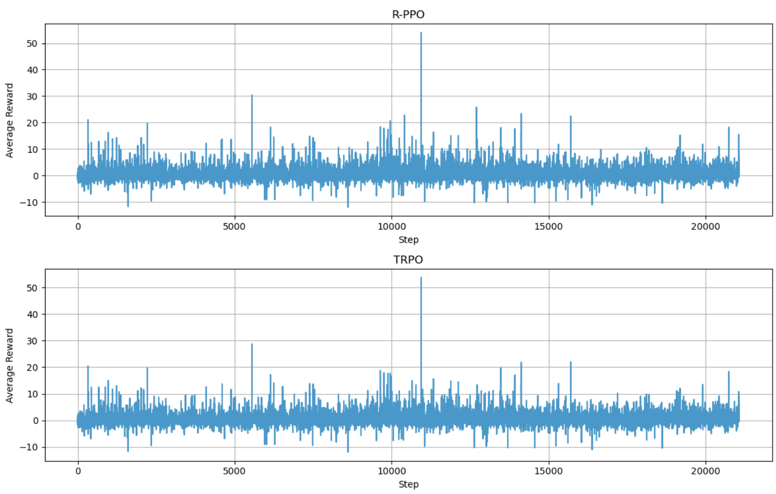

Figure 4.8: Average Rewards for: R-PPO and TRPO

Chapter 5

Conclusions and Future Work

The present research was set out to answer the research question, "What is the optimal discount range that encourages consumers to purchase perishables products and maximize retailer profitability?" through the application of reinforcement learning algorithms, including TRPO, PPO, R-PPO, A2C, and a benchmark model with a fixed 30% discount for comparison. The research evaluated the impact of different discount strategies on sales, profitability, and waste.

The findings reveal that while algorithms like TRPO successfully maximize sales volumes and profitability, this strategy comes at the cost of significantly higher waste and associated costs. Contrarily, PPO, due to extreme discounting, has increased unit sales but minimized sales volumes and profitability, resulting in lower waste and associated costs. TRPO is the preferred algorithm for scenarios where profitability is the goal.

For the raised research question, the results suggest that the optimal discount range that encourages consumers' purchases while balancing sales, profitability, and, in some cases, waste reduction lies between 30% and 60%. Discounts within this range generate substantial sales while minimizing the risk of over-discounting, which can lead to excess waste and reduced profitability.

It is expected that the insights from this study can have practical implications for retailers, especially for those managing perishable goods. By identifying a discount range that maximizes both sales and sustainability, retailers can adopt a discounting strategy that encourages customers to purchase without leading to unsustainable waste levels. Therefore, adopting these discounts (30%-60%) can help retailers maintain profitability by balancing inventory turnover and waste reduction.

5.1 Future Developments

Building on these findings, while it's expected that this research provides some insights into future works, some limitations should be acknowledged and explored further in future works.

- **Consumer Demand:** One essential improvement area is modeling customer demand. The current model simplifies consumer behavior by linking demand for discounted products to the level of discounts and for non-discounted products to their remaining shelf life. This approach may not fully capture the complexity of real-world purchasing decisions and how consumers behave in diverse retail environments. As a result, the model may not simulate customer demand with sufficient accuracy.

Future work should incorporate a more realistic demand model that captures the nuances of consumer behavior. This model should account for how customers respond to factors like brand loyalty, product freshness, and competitor pricing, as well as the availability of both discounted and non-discounted products. These improvements would allow the agent to adjust discounts, sales patterns, and restocking strategies more effectively, aligning them with real-world customer preferences.

- **Demand Forecasting:** Demand estimation can be enhanced by integrating more sophisticated models (e.g., time-series forecasting or machine learning-based demand prediction) that can help to improve the balance between inventory levels and consumer demand. This can lead to more accurate restocking decisions and better pricing strategies, reducing the reliance on discounts and limiting waste.
- **Reward Function:** The existing reward function may need to be refined to set a more significant focus on long-term profitability. Future work can focus on adjusting the reward structure to make profit margins, inventory holding costs, and the cost of unsold goods, more importantly, motivating the agent to optimize for long-term profitability.
- **Longer Training Periods and Algorithm Improvements:** Extending the training periods for the reinforcement learning agent and exploring more advanced algorithms, such as hybrid approaches combining RL with supervised learning or heuristics-based decision-making, could improve their ability to learn optimal long-term policies. Balancing sales, waste, and discount levels could lead to better profitability.
- **Incorporating Additional Constraints:** Future models could be enhanced by including additional real-world constraints such as supplier lead times, inventory storage costs, and shelf space limitations, menu costs, which arise from the operational burden of frequently changing prices. Accounting for these factors would create a more realistic environment and allow the RL algorithms to optimize performance more comprehensively.

Future work could address these areas and enhance the RL environment's ability to optimize retail operations, particularly in the challenging domain of perishable product management.

References

Godwyll Aikins, Sagar Jagtap, and Kim-Doang Nguyen. A Robust Strategy for UAV Autonomous Landing on a Moving Platform under Partial Observability. *Drones*, 8(6):232, June 2024. ISSN 2504-446X. doi: 10.3390/drones8060232. URL https://www.mdpi.com/2504-446X/8/6/232. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

Monica Alonso, Hortensia Amaris, David Martin, and Arturo de la Escalera. Proximal Policy Optimization for Energy Management of Electric Vehicles and PV Storage Units. *Energies*, 16 (15):5689, January 2023. ISSN 1996-1073. doi: 10.3390/en16155689. URL https://www.mdpi.com/1996-1073/16/15/5689. Number: 15 Publisher: Multidisciplinary Digital Publishing Institute.

Jocelyn M. Boiteau and Prabhu Pingali. Can we agree on a food loss and waste definition? an assessment of definitional elements for a globally applicable framework. *Global Food Security*, 37:100677, 2023. ISSN 2211-9124. doi: https://doi.org/10.1016/j.gfs.2023.100677. URL https://www.sciencedirect.com/science/article/pii/S221191242300007X.

M. E. Buisman, R. Haijema, and J. M. Bloemhof-Ruwaard. Discounting and dynamic shelf life to reduce fresh food waste at retailers. *International Journal of Production Economics*, 209: 274–284, March 2019. ISSN 0925-5273. doi: 10.1016/j.ijpe.2017.07.016. URL https://www.sciencedirect.com/science/article/pii/S0925527317302335.

Hsiu-Hua Chang and Jing-Wen Su. Sustainable consumption in Taiwan retailing: The impact of product features and price promotion on purchase behaviors toward expiring products. *Food Quality and Preference*, 96:104452, March 2022. ISSN 0950-3293. doi: 10.1016/j.foodqual.2021.104452. URL https://www.sciencedirect.com/science/article/pii/S0950329321003347.

Jing Chen, Ming Dong, Ying Rong, and Liang Yang. Dynamic pricing for deteriorating products with menu cost. *Omega*, 75:13–26, March 2018a. ISSN 0305-0483. doi: 10.1016/j.omega.2017.02.001. URL https://www.sciencedirect.com/science/article/pii/S030504831730172X.

Sheng-Chih Chen, Jie Min, Jinn-Tsair Teng, and Fuan Li. Inventory and shelf-space optimization for fresh produce with expiration date under freshness-and-stock-dependent demand rate. *Journal of the Operational Research Society*, 67(6), June 2016. ISSN 0160-5682. doi: 10.1057/jors.2015.100. URL https://doi.org/10.1057/jors.2015.100.

Wenchong Chen, Hongwei Liu, and Dan Xu. Dynamic Pricing Strategies for Perishable Product in a Competitive Multi-Agent Retailers Market. *Journal of Artificial Societies and Social Simulation*, 21(2):12, 2018. ISSN 1460-7425.

Wenchong Chen, Hongwei Liu, and Dan Xu. Dynamic Pricing Strategies for Perishable Product in a Competitive Multi-Agent Retailers Market. *Journal of Artificial Societies and Social Simulation*, 21(2):12, 2018b. ISSN 1460-7425.

Jaekwon Chung. Effective Pricing of Perishables for a More Sustainable Retail Food Market. *Sustainability*, 11(17):4762, January 2019. ISSN 2071-1050. doi: 10.3390/su11174762. URL https://www.mdpi.com/2071-1050/11/17/4762. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

Jaekwon Chung and Dong Li. A simulation of the impacts of dynamic price management for perishable foods on retailer performance in the presence of need-driven purchasing consumers. *Journal of the Operational Research Society*, 65(8):1177–1188, August 2014. ISSN 0160-5682. doi: 10.1057/jors.2013.63. URL https://doi.org/10.1057/jors.2013.63. Publisher: Taylor & Francis _eprint: https://doi.org/10.1057/jors.2013.63.

Ycaro Dantas, Pedro Enrique Iturria-Rivera, Hao Zhou, Majid Bavand, Medhat Elsayed, Raimundas Gaigalas, and Melike Erol-Kantarci. Beam Selection for Energy-Efficient mmWave Network Using Advantage Actor Critic Learning. In *ICC 2023 - IEEE International Conference on Communications*, pages 5285–5290, May 2023. doi: 10.1109/ICC45041.2023.10279804. URL https://ieeexplore.ieee.org/document/10279804. ISSN: 1938-1883.

FAO. *The State of Food Security and Nutrition in the World 2023: Urbanization, agrifood systems transformation and healthy diets across the rural-urban continuum.* FAO, IFAD, UNICEF, WFP, WHO, 2023. ISBN 978-92-5-137226-5. doi: 10.4060/cc3017en.

Food Security Information Network. Global report on food crises 2023. Available at https://www.wfp.org/publications/global-report-food-crises-2023, May 2023.

Adrienne Hall-Phillips and Purvi Shah. Unclarity confusion and expiration date labels in the United States: A consumer perspective. *Journal of Retailing and Consumer Services*, 35:118–126, March 2017. ISSN 0969-6989. doi: 10.1016/j.jretconser.2016.12.007. URL https://www.sciencedirect.com/science/article/pii/S0969698916302120.

Talha Harcar and Fahri Karakaya. A cross-cultural exploration of attitudes toward product expiration dates. *Psychology & Marketing*, 22(4):353–371, 2005. doi: https://doi.org/10.1002/mar.20063. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mar.20063.

on Food Security and Nutrition HLPE. Food losses and waste in the context of sustainable food systems. *Committee on World Food Security (CFS)*, 2014. URL https://www.fao.org/sustainable-food-value-chains/library/details/en/c/266229/.

Izabela Karolina Horoś and Tonia Ruppenthal. Avoidance of Food Waste from a Grocery Retail Store Owner's Perspective. *Sustainability*, 13(2):550, January 2021. ISSN 2071-1050. doi: 10.3390/su13020550. URL https://www.mdpi.com/2071-1050/13/2/550. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Wen-Shin Huang, Hung-Yu Kuo, Shi-Yuan Tung, and Han-Shen Chen. Assessing Consumer Preferences for Suboptimal Food: Application of a Choice Experiment in Citrus Fruit Retail. *Foods*, 10(1):15, January 2021. ISSN 2304-8158. doi: 10.3390/foods10010015. URL https://www.mdpi.com/2304-8158/10/1/15. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

INESC TEC. On incorporating consumer behaviour into the supply chain planning of fresh products. Available at https://www.inesctec.pt/en/projects/befresh, 2024.

Yasanur Kayikci, Sercan Demir, Sachin K. Mangla, Nachiappan Subramanian, and Basar Koc. Data-driven optimal dynamic pricing strategy for reducing perishable food waste at retailers. *Journal of Cleaner Production*, 344:131068, April 2022. ISSN 0959-6526. doi: 10.1016/j.jclepro.2022.131068. URL https://www.sciencedirect.com/science/article/pii/S0959652622007016.

Faruk Anıl Konuk. The role of risk aversion and brand-related factors in predicting consumers' willingness to buy expiration date-based priced perishable food products. *Food Research International*, 112:312–318, October 2018. ISSN 0963-9969. doi: 10.1016/j.foodres.2018.06.009. URL https://www.sciencedirect.com/science/article/pii/S096399691830454X.

Guillaume Le Borgne, Lucie Sirieix, and Sandrine Costa. Perceived probability of food waste: Influence on consumer attitudes towards and choice of sales promotions. *Journal of Retailing and Consumer Services*, 42:11–21, May 2018. ISSN 0969-6989. doi: 10.1016/j.jretconser.2018.01.004. URL https://www.sciencedirect.com/science/article/pii/S0969698917301844.

Rongpeng Li, Chujie Wang, Zhifeng Zhao, Rongbin Guo, and Honggang Zhang. The LSTM-Based Advantage Actor-Critic Learning for Resource Management in Network Slicing With User Mobility. *IEEE Communications Letters*, 24(9):2005–2009, September 2020a. ISSN 1558-2558. doi: 10.1109/LCOMM.2020.3001227. URL https://ieeexplore.ieee.org/document/9112328. Conference Name: IEEE Communications Letters.

Tongzhe Li, Kent D. Messer, and Harry M. Kaiser. The impact of expiration dates labels on hedonic markets for perishable products. *Food Policy*, 93:101894, May 2020b. ISSN 0306-9192. doi: 10.1016/j.foodpol.2020.101894. URL https://www.sciencedirect.com/science/article/pii/S0306919220300968.

Yanzhi Li, Brenda Cheang, and Andrew Lim. Grocery perishables management. *Production and Operations Management*, 21(3):504–517, 2012. doi: https://doi.org/10.1111/j.1937-5956.2011.01288.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1937-5956.2011.01288.x.

Cinzia Muriana. A focus on the state of the art of food waste/losses issue and suggestions for future researches. *Waste Management*, 68:557–570, 2017. ISSN 0956-053X. doi: https://doi.org/10.1016/j.wasman.2017.06.047. URL https://www.sciencedirect.com/science/article/pii/S0956053X17304889.

Ioana Popescu and Yaozhong Wu. Dynamic Pricing Strategies with Reference Effects. *Operations Research*, 55(3):413–429, June 2007. ISSN 0030-364X. doi: 10.1287/opre.1070.0393. URL https://pubsonline.informs.org/doi/10.1287/opre.1070.0393. Publisher: INFORMS.

Rupal Rana and Fernando S. Oliveira. Dynamic pricing policies for interdependent perishable products or services using reinforcement learning. *Expert Systems with Applications*, 42(1):426–436, January 2015. ISSN 0957-4174. doi: 10.1016/j.eswa.2014.07.007. URL https://www.sciencedirect.com/science/article/pii/S095741741400400X.

ReFED. Food waste problem. Available at https://refed.org/food-waste/the-problem/, 2024.

Lena Riesenegger and Alexander Hübner. Reducing Food Waste at Retail Stores—An Explorative Study. *Sustainability*, 14(5), January 2022. ISSN 2071-1050. doi: 10.3390/su14052494. URL https://www.mdpi.com/2071-1050/14/5/2494. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Zheng Rongcai, Xie Hongwei, and Yuan Kexin. Autonomous collision avoidance system in a multi-ship environment based on proximal policy optimization method. *Ocean Engineering*, 272:113779, March 2023. ISSN 0029-8018. doi: 10.1016/j.oceaneng.2023.113779. URL https://www.sciencedirect.com/science/article/pii/S0029801823001634.

Prajwalita Saikia, Sonia Pala, Keshav Singh, Sandeep Kumar Singh, and Wan-Jen Huang. Proximal Policy Optimization for RIS-assisted Full Duplex 6G-V2X Communications. *IEEE Transactions on Intelligent Vehicles*, pages 1–16, 2023. ISSN 2379-8904. doi: 10.1109/TIV.2023.3275632. URL https://ieeexplore.ieee.org/document/10123981. Conference Name: IEEE Transactions on Intelligent Vehicles.

Robert Evan Sanders. Dynamic Pricing and Organic Waste Bans: A Study of Grocery Retailers' Incentives to Reduce Food Waste, 2023.

Michael Scholz and Roman-David Kulko. Dynamic pricing of perishable food as a sustainable business model. *British Food Journal*, 124(5):1609–1621, January 2021. ISSN 0007-070X. doi: 10.1108/BFJ-03-2021-0294. URL https://doi.org/10.1108/BFJ-03-2021-0294. Publisher: Emerald Publishing Limited.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust Region Policy Optimization, April 2017a. URL http://arxiv.org/abs/1502.05477. arXiv:1502.05477 [cs].

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, August 2017b. URL http://arxiv.org/abs/1707.06347. arXiv:1707.06347 [cs].

Manu Sharma, Sudhanshu Joshi, and Kannan Govindan. Overcoming barriers to implement digital technologies to achieve sustainable production and consumption in the food sector: A circular economy perspective. *Sustainable Production and Consumption*, 39, July 2023. ISSN 2352-5509. doi: 10.1016/j.spc.2023.04.002. URL https://www.sciencedirect.com/science/article/pii/S2352550923000751.

Keerthana Sivamayil, Elakkiya Rajasekar, Belqasem Aljafari, Srete Nikolovski, Subramaniyaswamy Vairavasundaram, and Indragandhi Vairavasundaram. A Systematic Study on Reinforcement Learning Based Applications. *Energies*, 16(3):1512, January 2023. ISSN 1996-1073. doi: 10.3390/en16031512. URL https://www.mdpi.com/1996-1073/16/3/1512. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

Christoph Teller, Christina Holweg, Gerald Reiner, and Herbert Kotzab. Retail store operations and food waste. *Journal of Cleaner Production*, 185, June 2018. ISSN 0959-6526. doi: 10.1016/j.jclepro.2018.02.280. URL https://www.sciencedirect.com/science/article/pii/S0959652618306127.

Michael Tsiros and Carrie M. Heilman. The Effect of Expiration Dates and Perceived Risk on Purchasing Behavior in Grocery Store Perishable Categories. *Journal of Marketing*, 69(2): 114–129, 2005. ISSN 0022-2429. URL https://www.jstor.org/stable/30162048. Publisher: American Marketing Association.

United Nations. Goal 2: Zero Hunger, 2024. URL https://www.un.org/sustainabledevelopment/hunger/.

United Nations Environment. Food waste index report. Available at http://www.unep.org/resources/report/unep-food-waste-index-report-2021, March 2021.

Q. Wu and D. Honhon. Don't waste that free lettuce! Impact of BOGOF promotions on retail profit and food waste, 2023. URL https://journals.sagepub.com/doi/epdf/10.1111/poms.13884.

Cenying Yang, Yihao Feng, and Andrew Whinston. Dynamic Pricing and Information Disclosure for Fresh Produce: An Artificial Intelligence Approach. *Production and Operations Management*, 31(1):155–171, January 2022. ISSN 1059-1478. doi: 10.1111/poms.13525. URL https://doi.org/10.1111/poms.13525. Publisher: SAGE Publications.

Shuai Zhao and Wenli Li. Blockchain-based traceability system adoption decision in the dual-channel perishable goods market under different pricing policies. *International Journal of Production Research*, 61(13):4548–4574, July 2023. ISSN 0020-7543. doi: 10.1080/00207543.2023.2168309. URL https://doi.org/10.1080/00207543.2023.2168309. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00207543.2023.2168309.

Jiangbo Zheng, Yanhong Gan, Ying Liang, Qingqing Jiang, and Jiatai Chang. Joint Strategy of Dynamic Ordering and Pricing for Competing Perishables with Q-Learning Algorithm. *Wireless Communications and Mobile Computing*, 2021:e6643195, March 2021. ISSN 1530-8669. doi: 10.1155/2021/6643195. URL https://www.hindawi.com/journals/wcmc/2021/6643195/. Publisher: Hindawi.