

Article

Product Recommendation with Price Personalization According to Customer's Willingness to Pay Using Deep Reinforcement Learning

Ali Mahdavian ¹, Hadi Moradi ^{1,*} and Behnam Bahrak ²¹ Department of Electrical and Computer Engineering, University of Tehran, Tehran 14179-35840, Iran; mahdavian.ali@ut.ac.ir² Tehran Institute for Advanced Studies, Khatam University, Tehran 19916-33357, Iran; b.bahrak@teias.institute

* Correspondence: moradih@ut.ac.ir

Abstract

Integrating recommendation systems with dynamic pricing strategies is essential for enhancing product sales and optimizing revenue in modern business. This study proposes a novel product recommendation model that uses Reinforcement Learning to tailor pricing strategies to customer purchase intentions. While traditional recommendation systems focus on identifying products customers prefer, they often neglect the critical factor of pricing. To improve effectiveness and increase conversion, it is crucial to personalize product prices according to the customer's willingness to pay (WTP). Businesses often use fixed-budget promotions to boost sales, emphasizing the importance of strategic pricing. Designing intelligent promotions requires recommending products aligned with customer preferences and setting prices reflecting their WTP, thus increasing the likelihood of purchase. This research advances existing recommendation systems by integrating dynamic pricing into the system's output, offering a significant innovation in business practice. However, this integration introduces technical complexities, which are addressed through a Markov Decision Process (MDP) framework and solved using Reinforcement Learning. Empirical evaluation using the Dunnhumby dataset shows promising results. Due to the lack of direct comparisons between combined product recommendation and pricing models, the outputs were simplified into two categories: purchase and non-purchase. This approach revealed significant improvements over comparable methods, demonstrating the model's efficacy.

Keywords: recommendationsystems; dynamic pricing; deep reinforcement learning; intelligent marketing promotions; machine learning in e-commerce; willingness to pay



Academic Editor: Vicente García-Díaz

Received: 6 September 2025

Revised: 23 October 2025

Accepted: 29 October 2025

Published: 5 November 2025

Citation: Mahdavian, A.; Moradi, H.; Bahrak, B. Product Recommendation with Price Personalization According to Customer's Willingness to Pay Using Deep Reinforcement Learning. *Algorithms* **2025**, *18*, 706. <https://doi.org/10.3390/a18110706>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The rapid growth of the Internet, the Fourth Industrial Revolution, and advances in Artificial Intelligence (AI) have made business process automation essential for survival. In particular, the retail and e-commerce sectors rely heavily on data, with customer information and purchase histories representing key business assets. Among the most effective strategies for engaging customers in these industries are intelligent and data-driven promotions.

Retaining existing customers is considerably more cost-effective than acquiring new ones, with studies suggesting that acquiring a new customer can cost five to seven times more than retaining an existing one [1–3]. Consequently, intelligent and personalized promotions that strengthen customer loyalty are vital. These initiatives not only enhance

attachment to brands and organizations but also improve overall profitability [4]. Automating promotions through AI-based recommendation systems that consider customer interests, preferences, and purchase behavior has therefore become a strategic necessity.

AI's influence on marketing is transforming how businesses understand and engage with customers. AI-driven systems not only automate and optimize marketing processes but also provide insights into behavioral patterns, enabling more informed decisions. For example, AI-powered marketing platforms personalize promotions by analyzing large volumes of customer data and adapting recommendations in real time. This approach increases customer satisfaction, optimizes resource allocation, and enhances conversion rates [5].

Personalization is a key driver of customer satisfaction and loyalty, forming the basis for advanced strategies such as Dynamic Pricing (DP) [6]. DP has been extensively applied in e-commerce to optimize discounts and manage inventory [7]. In highly competitive markets, these strategies often require sophisticated modeling of spatial and temporal dynamics. For instance, Kim et al. [8] proposed a Spatial–Temporal Attention-based Dynamic Graph Convolutional Network (STAD-GCN) for predicting retail gasoline prices by integrating spatial relationships and temporal variations.

Empirical studies indicate that personalized strategies enhance both satisfaction and loyalty by addressing rational and emotional aspects of customer decision-making [9]. Such approaches strengthen long-term relationships, encourage repeat purchases, and foster positive word-of-mouth communication. Moreover, leveraging behavioral data allows companies to continuously refine personalization strategies to remain competitive and profitable in dynamic market environments.

Recommendation systems (RSs) are among the most successful applications of machine learning [10], widely adopted across domains such as e-commerce, online streaming, education, and social media [11]. In e-commerce, RSs help identify and recommend products aligned with users' preferences and behaviors, assist with product ranking during searches, and propose similar or complementary items. These systems personalize the shopping experience and significantly improve decision-making efficiency.

However, traditional RS typically focus only on identifying products that users may like, neglecting the critical influence of price. This omission can lead to suboptimal outcomes, as recommended products may fall outside a customer's preferred price range or willingness to pay (WTP), reducing the likelihood of purchase and overall profitability. To optimize effectiveness, RS must integrate pricing factors so that recommendations reflect both customer preferences and economic constraints. When profit and discount considerations are ignored, the RS remains incomplete and suboptimal [12]. Traditional systems tend to optimize metrics such as click-through rates or ranking quality but often overlook DP factors that directly affect sales conversion and revenue [13].

As illustrated in Figure 1, price personalization allows a seller to offer a product at a discounted price (price B) to attract additional buyers who might not have purchased at the original price (price A). While customers willing to pay price A continue to generate revenue, the inclusion of price B results in additional sales and higher overall profit [14]. Consequently, personalized pricing can substantially enhance business revenue.

Incorporating price personalization into RS is therefore a crucial objective of this research. Recommending products at prices aligned with each customer's WTP can significantly strengthen marketing strategies and improve promotional efficiency. This study introduces a reinforcement learning (RL)-based framework that jointly models product recommendations and price optimization. The proposed price-based promotion approach aims to increase customer retention and profitability by aligning product offerings with customers' purchase propensities.

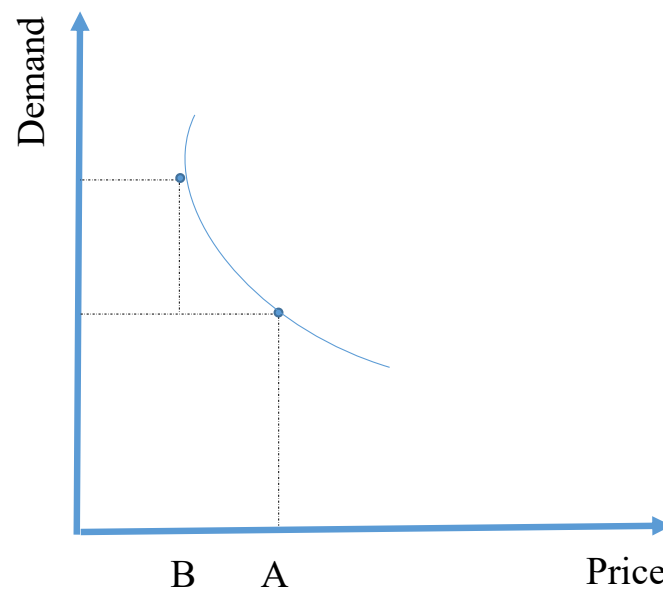


Figure 1. Increase in product sales at lower prices.

In traditional promotional campaigns, businesses often allocate a fixed budget to increase sales through uniform discounts—for example, offering 20% off all products. This method disregards individual differences in WTP and is suboptimal from both business and customer perspectives. Customers who are willing to pay the full price may become conditioned to expect discounts, reducing future margins, whereas those requiring higher discounts may remain unmotivated to purchase.

A more effective strategy involves tailoring discounts and prices to each customer's WTP. For instance, offering a product at full price to a customer ready to pay it, and the same product at a 40% discount to another customer who requires that incentive, results in successful purchases by both. This targeted approach maximizes both customer satisfaction and efficient budget utilization.

As illustrated in Figure 2, offering personalized prices aligned with individual customer characteristics supports both marketing personalization and operational optimization objectives.

In marketing theory, the concept of the marketing mix—commonly referred to as the 4Ps (Product, Price, Promotion, and Place)—is fundamental. To maximize customer impact, it is essential to deliver the right product at the right price, supported by effective promotional strategies and efficient distribution channels. This interdependence highlights the critical importance of integrating promotion and pricing strategies within the overall marketing framework.

As shown in Figure 3, effective pricing should be tailored to the specific characteristics of each customer. Offering a product at different prices to different customers is not only a common commercial practice but also aligns with personalization and optimization principles central to modern marketing.

Similarly, Figure 4 illustrates how a customer may not be inclined to purchase a product at full price but may complete the transaction if a discount is applied according to their profile or behavioral characteristics.

To design a model capable of recommending both products and prices aligned with customers' purchase propensities, various statistical and machine learning techniques can be employed. After reviewing prior research and evaluating different methods, RL emerged as a particularly effective approach for this purpose.

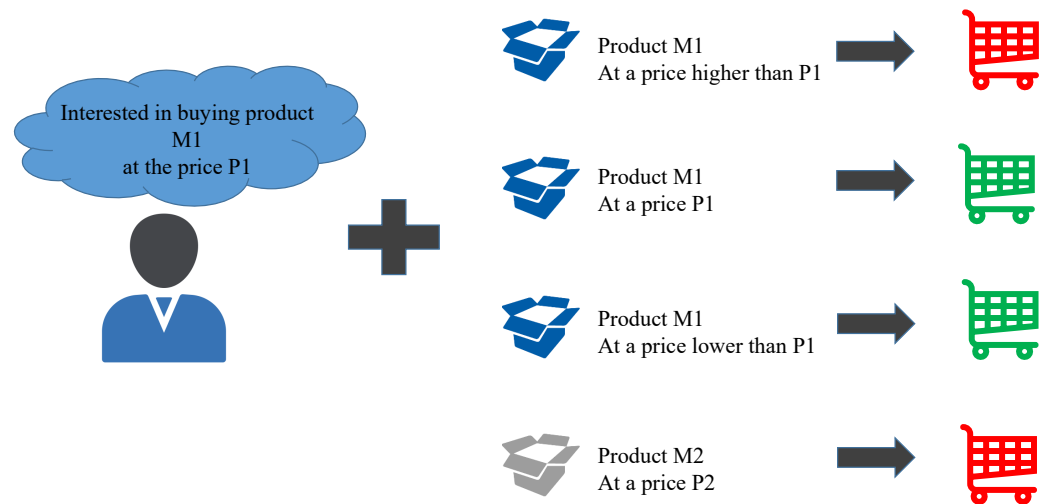


Figure 2. Smart promotion design tailored to the customer's purchase intention for a specific product.

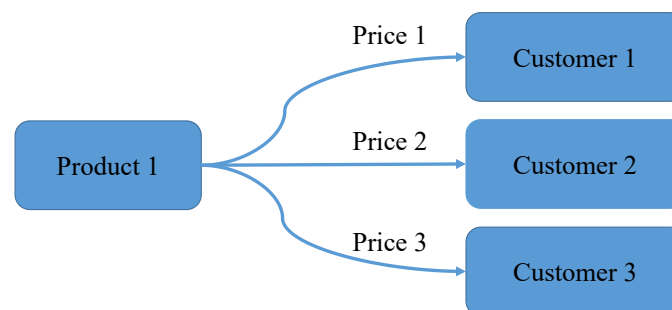


Figure 3. Personalizing product prices for customers.

RL offers significant advantages for RS by optimizing long-term user engagement and satisfaction. Traditional RS methods, such as collaborative or content-based filtering, typically optimize for short-term objectives (e.g., clicks or immediate purchases). In contrast, RL maximizes cumulative rewards over time, providing several benefits.

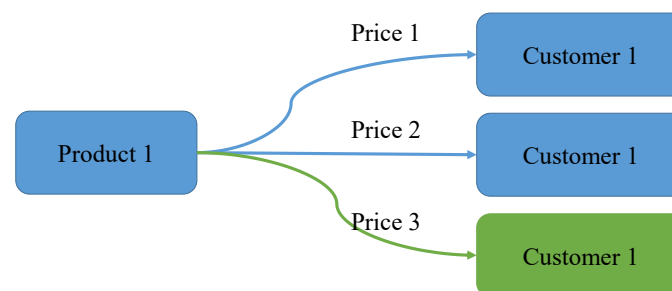


Figure 4. Customer's purchase intention for a specific product.

First, RL supports sequential decision-making by considering the temporal order of user interactions, allowing the system to learn from past behaviors and anticipate future preferences [15]. Second, it adapts dynamically to real-time feedback, continuously improving its performance as new data are observed [16]. Third, RL enhances personalization by optimizing for long-term rewards, thereby tailoring recommendations more effectively and fostering user loyalty [17]. Additionally, it balances exploration and exploitation—discovering new items while leveraging known user preferences [15]—and enables holistic optimization of multiple metrics such as retention, lifetime value, and engagement [15].

RL has also demonstrated strong performance in sequential decision-making through prioritized experience replay, which accelerates convergence and improves profitability by

focusing on high-impact interactions [18]. In summary, RL enables RS to evolve from static, one-step predictors into adaptive systems that optimize engagement and value across time.

Deep Reinforcement Learning (DRL) extends these advantages by combining RL with the representational power of deep neural networks. DRL is particularly well suited for recommendation and pricing applications due to the following capabilities:

- **Handling High-Dimensional Data:** DRL uses deep neural networks to manage large and complex state–action spaces, which are common in RS owing to the vast number of potential user–item interactions. This allows the model to capture intricate patterns and nonlinear relationships in user data [16].
- **Robustness to Sparse Data:** DRL methods are resilient to the data sparsity that often affects collaborative filtering techniques. Their ability to generalize from limited data makes them particularly valuable for addressing cold-start problems involving new users or products [15].

Accordingly, this research proposes a DRL-based model that recommends products together with optimal prices tailored to customers’ purchase propensities. This integration supports the design of intelligent, price-aware promotions that maximize both customer satisfaction and business profitability.

After formulating the problem as a Markov Decision Process (MDP), the Deep Q-Network (DQN) architecture was employed to solve it using DRL techniques.

Recent developments in DRL have expanded its application to large-scale recommendation and DP scenarios. For example, Wang et al. [19] introduced a deep Q-learning-based pricing framework that adapts dynamically to market fluctuations, achieving significant revenue gains. Similarly, Kavosi et al. [20] applied DRL to joint inventory and pricing optimization in retail, demonstrating scalability across thousands of products. In the domain of personalized recommendations, Tanveer et al. [21] proposed a graph-attention-augmented DRL model (PGA-DRL) that captures complex relationships between users and items to improve personalization. Together, these studies underscore how DRL enables scalable, adaptive, and context-aware pricing strategies—supporting the motivation for our DQN-based framework that integrates product recommendation and pricing in a unified model.

In this study, the proposed framework was implemented using the Dunnhumby sales dataset, which contains detailed records of product sales, prices, and applied discounts. Model evaluation on the test dataset yielded strong results in the multi-class classification task, with a Macro Average Precision of 0.80, Macro Average Recall of 0.82, Macro F-score of 0.81, Micro F-score of 0.85, Macro Averaged AUC of 0.87, NDCG@5 of 0.89 and Weighted Multi-class Accuracy of 0.80. These results confirm the effectiveness of the proposed reinforcement learning framework in accurately recommending both products and corresponding prices according to customers’ purchase tendencies.

Building upon these advances, this study contributes to the literature by introducing an integrated framework that unifies personalized product recommendation and price optimization within a single RL environment. The model extends existing RS research by explicitly incorporating customers’ willingness to pay, thereby bridging the gap between DP and intelligent recommendation. By embedding pricing into the recommendation output, the model enhances the scope, accuracy, and practical relevance of traditional RS. Its applicability spans various industries—particularly retail and e-commerce—where product diversity, price sensitivity, and customer heterogeneity are prominent.

Contributions and Innovations.

This research addresses a novel and practically significant problem by integrating personalized product recommendation and DP within a unified DRL framework. Unlike

traditional RSs that optimize short-term engagement or rely on static pricing models, the proposed approach jointly models customer preferences and WTP, enabling adaptive promotional strategies. The framework leverages recent advances in DRL to design intelligent, price-based promotions that simultaneously optimize customer satisfaction and long-term profitability. Additionally, it introduces problem-specific evaluation metrics that capture both financial relevance and predictive performance. Overall, this study combines theoretical rigor with practical application, contributing to the expanding body of DRL-based marketing research while offering measurable improvements in operational efficiency and profit optimization.

2. Related Work

2.1. RS

RSs are one of the innovative solutions designed to meet the needs of various industries, particularly in e-commerce and retail. These systems utilize customer behavior, customer information, and product information to extract customer preferences and suggest products that are likely to be purchased or liked by users [22]. RSs not only personalize the shopping experience but also enhance decision-making efficiency by reducing cognitive costs and improving decision quality [23]. Studies have shown that the implementation of RS in the e-commerce industry can lead to a 30–35% increase in the number of purchases (<https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers>, accessed on 9 September 2025).

2.1.1. Types of Data Sources Used in RS

RSs require an understanding of both customer behavior and product attributes, necessitating the use of diverse data sources to comprehend customers and products. The most significant data source is the interaction data between the customer and the product. This interaction is categorized into explicit and implicit feedback. Explicit interactions are the cornerstone of RS, reflecting direct customer opinions on a product. Direct customer ratings are among the most crucial explicit interactions, indicating direct customer satisfaction with a product. Extracting the semantic content of customer reviews on products and recommending products to friends are other forms of explicit interactions.

Conversely, implicit interactions refer to those that do not necessarily convey direct customer satisfaction but imply indirect satisfaction or a relationship between the customer and the product. Examples of implicit interactions include purchasing a product, viewing a product page, adding a product to a shopping cart, and clicking on a product link. Explicit interactions are usually fewer in number and include both positive and negative feedback, where customers explicitly express their likes or dislikes for a product. In contrast, implicit interactions are more numerous and mostly include positive feedback, as reliable data on negative implicit feedback is generally unavailable [24].

Explicit interactions typically have less error compared to implicit interactions because customers explicitly state their opinions [24]. However, implicit interactions require inference from available data, which can introduce errors due to insufficient information. For instance, merely purchasing a product does not necessarily indicate customer satisfaction, unlike explicit ratings that clearly convey satisfaction levels. Other examples of implicit interaction errors include interpreting page views as customer interest or adding products to the cart out of curiosity rather than purchase intent.

Studies [24,25] have utilized weighted combinations of explicit and implicit interactions to create a final score, referring to these as customer behavioral data. Additionally, customer demographic data such as age, gender, education level, and income, along with product data like category, subcategory, and brand, are also valuable data sources.

The study by Bai et al. [26] proposes a novel Gaussian-based optimization criterion to improve recommender systems using implicit feedback. This approach addresses the challenge of assigning uniform confidence weights to all user-item interactions by modeling user preference confidence as a learnable variable.

Pujahari and Sisodia [27] proposed a novel framework using Markov Random Fields to model dynamic user preferences in sparse and time-variant datasets. By integrating user and item features with temporal information, this approach enhanced recommendation quality and addressed cold-start challenges.

While implicit interactions are inherently noisier than explicit feedback, their predominance in real-world retail datasets makes them a more realistic foundation for scalable modeling. Moreover, accurately inferring customer WTP requires access to both behavioral interaction data and detailed price or discount information—a combination rarely available in public datasets. In this study, despite extensive dataset exploration, no publicly accessible source was found that included explicit user ratings alongside price and discount variables. Consequently, focusing on implicit data represents a necessary and pragmatic compromise between data availability and model generalizability. To mitigate the inherent uncertainty of implicit signals, our framework incorporates a diverse set of behavioral and transactional features, enabling the reinforcement learning model to infer WTP through aggregated behavioral patterns rather than isolated actions.

Another crucial data source in RS is related to cognitive sciences, personality, and emotions. Personality traits, such as extraversion, agreeableness, and neuroticism, which are stable over time, can be derived using algorithms and questionnaires. Emotions, which are more variable and can change over time and conditions, such as sadness, anger, or happiness, can be detected through cameras and physiological sensors [28,29].

Contextual data indicating the customer's environmental conditions can also be used in RS [30]. Examples of contextual features include geographical location and time. Various studies [31–33] have utilized these contextual features to enhance RS. Another data source is session data, which contains information about the current interaction between the customer and the business, without historical data [10].

Product prices are also used as a data source in numerous studies. Price-aware RSs use price as an input for recommending products to customers but do not incorporate it into the output.

2.1.2. RS Methods

There are three traditional categories of methods used in RS: collaborative filtering (CF), content-based filtering (CBF), and hybrid methods [34].

CF: CF recommends products to users based on the preferences of other users. One key advantage of this method is that it does not require information about the products themselves. However, a significant drawback is that it struggles to provide suitable recommendations for new customers, for whom there is no historical interaction data. Additionally, it tends to favor popular products with many interactions, while products with fewer interactions are less likely to be recommended.

Within CF, there are two main approaches: item-based and user-based.

- **Item-based CF:** This approach assumes that if two products receive similar high ratings, a user who likes one of them is likely to also like the other.
- **User-based CF:** This approach measures the similarity between users based on their characteristics. If two users are found to be similar, a product that one user likes is likely to be liked by the other as well.

CBF: CBF operates based on the similarity between product features. If a user likes a particular product, they are likely to be interested in similar products. Advantages of

this method include its ability to provide recommendations even when a user has few interactions. However, it requires descriptive data for products and can be slow if there are many products.

Hybrid Methods: Hybrid methods combine CF and CBF to leverage the strengths of both approaches. These methods aim to provide more accurate recommendations by overcoming the limitations of each individual method.

Additionally, there are classification-based collaborative filtering methods that predict whether a user will purchase a product based on historical data, contextual data, and user and product features. Another method is popularity-based recommendation, which suggests more popular products. Correlation-based recommendation calculates the correlation coefficient between products, suggesting that if a product chosen by a customer has a high correlation with another product, it implies a strong positive linear relationship between them.

2.1.3. Types of RS by Output

Traditionally, RSs aim to suggest individual products that customers might be interested in. However, other categories of systems exist that focus on recommending groups or bundles of products to customers. This approach has been explored in various studies [35–38]. Additionally, research [39] discusses product recommendations to groups of customers, known as group recommendations. Liu et al. [38] propose a multi-view graph contrastive learning model to enhance bundle recommendations by capturing complex user-item interactions across multiple views. This approach uses graph neural networks to integrate diverse interactions, producing cohesive bundles that align more closely with user interests. The model's multi-view structure significantly improves bundle recommendation quality over traditional methods.

Thus, the most common output of RS, as referenced in most studies and applications, is the recommendation of individual products to users. This output can also include pricing, ensuring that the price aligns with the customer's WTP for the product. However, this aspect has received less attention in research. The significance of this method and the proposed approach to address it are further discussed.

2.2. Price in RS

2.2.1. Promotion Analysis

AI-powered marketing systems are reshaping the way businesses design and implement promotional strategies. By leveraging large-scale datasets and advanced analytics, companies can tailor promotions to individual customer preferences and behaviors, delivering highly relevant and timely offers that drive engagement. The ability of AI to analyze historical behavior and predict future trends enables a level of personalization previously unattainable, thereby enhancing the overall customer experience [5].

Promoting the right product to the right customer at the right price—aligned with the customer's WTP—is essential for effective marketing. Such alignment not only strengthens customer retention and loyalty, leading to higher revenue and profitability, but also enhances satisfaction through meaningful and personalized interactions.

Beyond product quality, the quality of a company's interactions with its customers plays a decisive role in shaping satisfaction and loyalty. Intelligent promotions allow businesses to act proactively, recognizing and valuing the individuality of their customers through customized offers. These personalized experiences can evoke positive emotional responses, differentiating the business from competitors and creating memorable connections. As a result, intelligent promotions represent one of the most powerful tools for fostering emotional engagement and building lasting customer relationships.

The success of intelligent promotion strategies depends not only on the accuracy of product recommendations, as discussed in the context of RS, but also on the timing and pricing of these offers. Timing is a critical factor—elements such as product shelf life, seasonal demand, or promotional cycles significantly influence effectiveness [40]. Equally important is the pricing dimension, which determines how well a promotion resonates with customer expectations and perceived value.

2.2.2. The Importance of Pricing in Promotions

In line with the effectiveness of RS, their applications have evolved from information retrieval towards automated marketing tools. The primary goal of these marketing tools is to have a direct positive impact on customer's purchase decisions, and it is well known that customer buying behavior is heavily influenced by price [25]. Most research in the field of RS focuses on optimizing quantitative metrics based on historical data, such as accuracy and coverage. However, there is a gap between academic research and industry practices, where metrics like revenue and profit are more prominent [41]. Traditional RSs have primarily focused on optimizing user satisfaction metrics such as click-through rates or ranking quality. However, value-aware recommender systems (VARs) shift the focus to economic objectives, balancing stakeholder interests such as revenue maximization and customer retention [13]. Modeling and predicting prices in competitive markets, particularly in retail, play a significant role in strategic decision-making. Pricing not only influences customer purchase decisions but also helps businesses optimize market share and profitability by considering dynamic spatial and temporal factors [8].

RS must not only focus on accuracy but also incorporate user-centric features that address transparency and fairness, thereby fostering trust and long-term engagement [42].

Study [43] shows that strategic pricing, such as dynamic and personalized pricing, enhances the effectiveness of recommendations by aligning product prices with customer expectations and WTP.

Price is a crucial element that determines whether a product will ultimately be purchased [44]. The pricing of products has a direct impact on the revenue and profit of businesses. Additionally, the price presented to the customer, whether the full price or a discounted price, affects their buying behavior and satisfaction. Therefore, selecting the appropriate price for products according to customer's willingness to purchase is essential.

In understanding the role of pricing within promotions, it is crucial to recognize both the rational and emotional factors that influence customer's WTP. According to [45], WTP serves as a measure of the rational, task-oriented value that users assign to information objects, while Experienced Utility reflects the emotional response to these objects. This dual approach suggests that an effective pricing strategy must consider not only the logical assessment of product value but also the emotional appeal, as pricing that resonates emotionally with customers can enhance the perceived value and appeal of promotional offers. By aligning promotional prices with both rational expectations and emotional engagement, businesses can potentially increase customer satisfaction and drive purchase decisions more effectively.

Unlike product attributes such as the manufacturer, which influence a customer's interest in a product, price directly affects the customer's willingness to buy that product. In other words, price and other features in the customer's purchase decision process are orthogonal to each other, and in most cases, the customer will make a purchase when both the price and product features are acceptable [44]. Thus, price is highly significant.

AI's role in pricing is crucial for creating customer-centric pricing models. Through analyzing historical pricing data, AI can dynamically adjust prices in real-time based on market conditions and customer behavior. This strategic use of AI ensures that businesses

remain competitive and maximize profitability, aligning product pricing with customer expectations and purchasing power. Such approaches not only enhance the success of promotions but also optimize pricing strategies to match real-time customer preferences, ultimately leading to increased sales and customer satisfaction [5].

The price feature can be considered from two perspectives: product characteristics and user preferences, as input to RS [46]. Study [46] evaluates the effectiveness of including price from either or both perspectives in RS positively.

Overall, there is limited research on incorporating price into RS [25]. This scarcity may be due to the challenging nature of the issue, as evidenced by the dimensions and challenges encountered in investigations, including dataset, modeling, algorithm, and evaluation methods. Marketing studies clearly indicate the importance of price as a critical factor influencing customer behavior and product sales [47]. Price is also crucial in promotions; merely reducing the price may not be beneficial, as it might not lead to customer satisfaction and could result in suboptimal business revenue and profit. Hence, determining the appropriate product price and offering suitable discounts are vital [48].

Recommending low-priced products to a customer who prefers to spend more on their purchases decreases business efficiency. Conversely, recommending high-priced products to a customer who tends to buy lower-priced items also reduces the effectiveness of the recommendation [25]. This issue is also true for price and discount levels.

Price personalization is a pricing scheme that allows sellers to adjust product prices based on individual users [49], also known as DP. Price personalization is a form of price differentiation, which is a pricing strategy where different prices are set for the same product. This differentiation can be based on various factors such as geographical location and demographic information like age, gender, etc. [14].

Study [12] emphasizes that if profit and discount are not considered in recommendations, the system is suboptimal. This is because a product might align with a customer's preferences, but its price may not be attractive enough. Therefore, the price must match the customer's WTP for the RS to be fully effective and for the purchase to occur.

However, defining the price is highly dependent on external constraints such as competitor pricing, production costs, and so on, and the recommended price by the RS must be approved by management. Therefore, including price in the output has this challenge and requires at least an additional step after the recommendation, where the proposed discount price is either approved or rejected by management [25].

Regarding the importance and impact of price, study [50] highlights the significance of discounts on customer buying behavior and proposes a discount-sensitive model. Study [51] also discusses the importance of considering price and profit in RS and incorporates past user preferences into the RS, proposing a price-sensitive model.

In study [52], customer discount rates are considered an important factor in the RS. Study [12] presents ideas for applying discounts to promoted products and suggesting non-discounted products alongside them, addressing issues such as which complementary, independent, or substitute products should be recommended together and at what prices.

The goal of study [44] is to develop an effective method for predicting customer purchase intentions with a focus on the price factor in RS. It uses price as an input feature for product recommendations to customers and compares its proposed method with basic methods that also consider price. This study highlights two problems of considering price in RS and proposes solutions for them. The first problem is the unknown preference and sensitivity of customers to product prices, implicitly reflected in the products they have purchased. The second problem is the extent to which the product price influences the customer's purchase intention, which is highly dependent on the product category. The customer's understanding and financial capability concerning a product are highly

dependent on its category. For the first problem, a transitive relationship between customer-product and product-price is modeled using a graph convolutional network. The main idea is to propagate the price effect on the customer through the product bridge. For the second problem, aggregating product categories in the propagation process and modeling potential binary interactions to predict customer-product interactions is suggested. Studies [25,43,46,48] have used price to better estimate user preferences.

In study [46], questions regarding the impact of price in RS and recommendations from unobserved categories are explored. This study proposes a metric called “price level” within product categories, expressed as a percentage range, indicating the price position of a product within that category. The method for calculating the price level for a product with a specific price within a category containing products with specific minimum and maximum prices is detailed in the following formula:

$$\text{Price Level} = \frac{\text{Product Price} - \text{Category Minimum Price}}{\text{Category Maximum Price} - \text{Category Minimum Price}}$$

In study [25], the main challenges in designing price-sensitive RS are described, showing the impact of price on the accuracy of RS and business efficiency. This study incorporates price alongside other inputs in the RS.

Study [53] focuses on maximizing the expected profit of the business owner using RS. Study [41] increases the accuracy of the RS by combining the profit of recommended products with customer preferences. This research suggests a profit-based evaluation metric for RS, paying more attention to products with higher profit in the recommendations.

Study [14] proposes price personalization, a pricing scheme to decide whether to offer a discount to a customer, and proposes RS with price personalization. This study uses user preference data for product recommendations. Additionally, patterns extracted from the previous step and historical purchase data are used for price adjustment through RL. Customers are categorized into three types based on historical purchase data: regular customers, discount customers, and indifferent customers.

WTP is defined as the highest acceptable price of a product for a customer [44], with the reservation price being the marketing term equivalent to WTP. Given the importance of price, the proposed RS in this study includes an additional feature where the product price is adjusted for the customer based on their characteristics and WTP.

2.3. RL

Unlike supervised or unsupervised learning, RL focuses on goal-directed learning, where an agent learns to maximize cumulative rewards through continuous interaction with its environment. Two defining characteristics distinguish RL from other learning paradigms: *guided trial-and-error exploration* and the optimization of *long-term cumulative rewards*. In RL, the agent observes the environment, takes actions, receives feedback in the form of rewards, and updates its policy to maximize future rewards [16]. The general structure of the RL framework is illustrated in Figure 5.

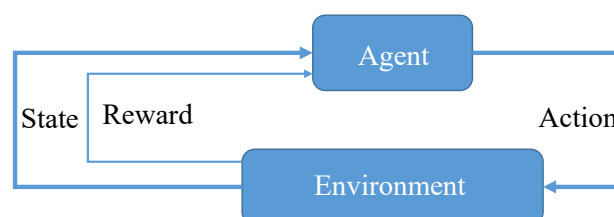


Figure 5. RL Schema.

For the agent to interact with the environment, the Markov Decision Process (MDP) framework must be followed [54]. This framework includes five components:

- **State Definition:** Describes the different situations or configurations in which the agent can be.
- **Action Definition:** Specifies the set of actions the agent can take in each state.
- **Transition Probability:** Defines the probability of moving from one state to another after taking a specific action.
- **Reward Definition:** Specifies the reward received after performing a particular action in a specific state.
- **Discount Factor:** A factor between 0 and 1 that determines the importance of future rewards compared to immediate rewards, affecting the cumulative reward calculation.

The discount factor determines how future rewards are weighted in the cumulative reward and how much immediate rewards at different times influence the discounted reward. This framework essentially models the world and environment in which the agent can interact, using guided trial and error to reach its goal, optimize its objective function, and select the best action with the highest long-term benefit based on each state.

The definition of this framework should ensure that the past interactions of the agent do not affect the prediction of rewards and subsequent states. Statistically, if the probability P of receiving a reward and transitioning to the next state when taking a specific action in the current state is considered, this probability should be independent of states, rewards, and actions before the current state and should always remain P .

The optimal policy that allows the agent to receive the highest rewards is a greedy policy, where in each state, the agent learns and takes the action with the highest reward. If the policy the agent uses to explore different states and rewards during the learning process is the same as the policy it optimizes, the learning process is called **on-policy**. However, in some cases, to reduce the agent's regret and improve the exploration process, two separate policies can be used: a behavior policy for the agent's interactions and a target policy for optimization. This process is known as **off-policy** learning.

RL methods are divided into two main categories: model-based and model-free approaches. In model-based methods, a predefined or trained model is used to learn the policy. In contrast, model-free methods directly learn the policy without having a model of the transition function [16].

Based on the agent's actions, RL methods can also be categorized into three types: value-based methods, policy search methods, and actor-critic methods [16]:

- **Value-Based Methods:** These are some of the fundamental RL methods where the optimal policy is derived by selecting actions that maximize the agent's value function.
- **Policy Search Methods:** These methods optimize the policy directly.
- **Actor-Critic Methods:** These methods combine value estimation and policy optimization by estimating the value function while simultaneously searching the policy space for the optimal policy. They leverage aspects of both previous methods to learn the value function and the policy [55].

Wang et al. [56] propose a novel multi-objective optimization (MOO) approach for permanent magnet machines using an improved soft actor-critic (SAC) algorithm. This study highlights the effectiveness of reinforcement learning in balancing conflicting objectives. In studies [57,58], RL has been utilized in RS.

Deep learning focuses on appropriately representing input features through multilayer networks of neurons and nonlinear transformations, capable of managing large numbers of features [55]. RL, as a continuous decision-making process, operates by maximizing the rewards received from the agent's interactions with the environment. In complex

applications, the high-dimensional nature of the problem can pose challenges for RL. Therefore, combining RL with deep learning, known as DRL, can address these high-dimensional issues. DRL adaptively selects an appropriate cooperation strategy to guide the evolution direction, which greatly improves the ability to balance convergence and diversity [59].

Various algorithms and methods under deep learning employ layers of neurons with different sequences, structures, and features for specific purposes. DRL has numerous applications, including computer vision, robot control, healthcare, and RS. It significantly enhances the capabilities of RL by adding unsupervised generative models, attention mechanisms, and memory to DRL networks, making agents smarter and more capable of performing tasks similar to human brain activities, such as planning and learning.

Research [60,61] has employed DRL in RS, demonstrating its potential and effectiveness in this domain.

DQN is a type of RL algorithm that combines Q-learning with deep neural networks to handle high-dimensional input spaces [62]. It was introduced by DeepMind in 2015 and has been successfully applied to a range of complex tasks, most famously in achieving human-level performance in Atari 2600 games [63].

Q-learning is a model-free RL algorithm that aims to learn the optimal action-selection policy using a Q-function, which estimates the expected reward of taking a certain action in a given state and following the optimal policy thereafter. The Q-function is updated iteratively using the Bellman equation:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left(r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right)$$

where:

- s and s' are the current and next states, respectively.
- a and a' are the current and next actions, respectively.
- r is the reward received after taking action a in state s .
- α is the learning rate.
- γ is the discount factor.

DQN leverages deep neural networks to approximate the Q-function, which allows it to handle complex state spaces like images, where traditional tabular methods would be infeasible.

The DQN algorithm incorporates several key components to improve stability and learning efficiency:

- **Experience Replay:** Instead of learning from each step of experience individually, DQN stores a finite number of past experiences in a replay buffer. During training, random samples from this buffer are used to update the Q-network, breaking the correlation between consecutive experiences and improving learning efficiency.
- **Target Network:** DQN employs two networks: the Q-network and the target Q-network. The Q-network is updated frequently, while the target Q-network is updated less frequently (typically every few thousand steps). This approach stabilizes learning by providing consistent target values for Q-learning updates. The target Q-value is computed as:

$$y = r + \gamma \max_{a'} Q_{\text{target}}(s', a')$$

The Q-network parameters are updated to minimize the difference between the Q-value and the target Q-value y .

DQN Algorithm Steps are presented in Algorithm 1, which outlines the procedure used for training the RL agent in this study.

Algorithm 1 Deep Q-Network (DQN) Algorithm

```

1: Initialization:
2: Initialize Q-network  $Q$  with random weights.
3: Initialize target Q-network  $Q_{\text{target}}$  with weights  $\theta_{\text{target}} = \theta$ .
4: Initialize replay buffer  $D$ .
5: for each episode do
6:   Set initial state  $s_0$ .
7:   for each time step do
8:     Select action  $a_t$  using an  $\epsilon$ -greedy policy based on  $Q$ .
9:     Execute  $a_t$ , observe reward  $r_t$  and next state  $s_{t+1}$ .
10:    Store  $(s_t, a_t, r_t, s_{t+1})$  in  $D$ .
11:    Sample random mini-batch of transitions  $(s, a, r, s')$  from  $D$ .
12:    Compute target Q-value for each transition:

```

$$y = r + \gamma \max_{a'} Q_{\text{target}}(s', a')$$

```

13:      Update  $Q$  by minimizing loss between predicted  $Q(s, a)$  and target  $y$ .
14:      Every  $C$  steps, set  $\theta_{\text{target}} \leftarrow \theta$ .
15:    end for
16: end for

```

By combining Q-learning with deep neural networks, DQN can efficiently learn optimal policies for complex tasks with high-dimensional state spaces. The use of experience replays and target networks are critical innovations that address the instability issues commonly associated with DRL [63].

Recent advances in DRL have considerably broadened its application scope beyond classical control problems to include DP, RS, and inventory optimization. Modern DRL models demonstrate improved scalability, robustness, and interpretability, making them highly suitable for complex data-driven decision-making environments.

In DP, Li et al. [64] proposed a Deep Q-Learning optimization framework for adaptive e-commerce pricing, enabling real-time adjustment to market fluctuations and outperforming static heuristic baselines. Similarly, Fraija et al. [65] applied DRL to demand response pricing, integrating supply and market constraints to optimize aggregator profit and customer comfort. Nomura et al. [66] used Proximal Policy Optimization (PPO) to optimize pricing and inventory policies for perishable goods, achieving over 60% faster computation and near-optimal profitability, while Kavoosi et al. [20] proposed a unified DRL framework combining inventory control and pricing in retail contexts.

In RS, Tanveer et al. [21] developed the PGA-DRL model, which incorporates progressive graph attention mechanisms to enhance relational learning between products and users. Wang et al. [67] introduced a hierarchical DRL recommendation algorithm that leverages multi-level attention mechanisms to model both short- and long-term user preferences. Rossiiev et al. [68] provided a comprehensive survey of RL-based RSs, highlighting scalability, interpretability, and hybrid model architectures. Gupta et al. [69] further explored multi-agent DRL approaches (MADDPG, MADQN, and QMIX) for interdependent pricing strategies in supply chain environments. Zhang et al. [70] conducted an updated survey on RL-based RSs, emphasizing integration with graph neural networks and large-scale industrial deployments.

One of the central challenges in applying reinforcement learning to real-world recommendation and pricing is scalability to large product catalogs and action spaces. Traditional RL methods struggle as the number of discrete actions (e.g. products $\times$ discount levels) grows, due to computational explosion in evaluating and updating Q-values. For example, Dulac-Arnold et al. address large discrete action spaces by embedding actions into a continuous representation and using approximate nearest-neighbor search to achieve sublinear

scaling in action lookups [71]. More recently, Liu et al. decompose recommendation actions into hyper-actions and effect-action steps to reduce the complexity of decision-making over vast item sets [72]. These methods illustrate how hierarchical, embedding-based, or decomposed action models can help RL systems scale.

2.4. Challenges and Research Gap

2.4.1. Data Availability and Interaction Types

RSs face several challenges related to data availability and interaction types. One major limitation is the scarcity of explicit feedback and the limited volume of implicit interaction data, as discussed in Section 2.1.1. Additional challenges include the lack of in-action data [32], the absence of negative feedback in implicit interactions, and the phenomenon of interest drift—the evolution of customer preferences and needs over time. These issues can be partially mitigated by incorporating historical behavioral data [10].

Negative customer sentiment resulting from exposure to unsuitable product promotions also poses a challenge [51]. According to Zhang et al. [38], RS often struggle with sparse data and complex user–item interaction patterns, which reduce their predictive accuracy and generalization ability. To address these limitations, Liu et al. [73] proposed the Interest Evolution-driven Gated Neighborhood (IEGN) model, which uses gated neighborhood aggregation to capture dynamic shifts in user interests. This approach enhances responsiveness and relevance even in data-sparse environments, ensuring that recommendations adapt effectively to changing user behavior.

2.4.2. Feature Space and Dynamic Nature

Another significant challenge in RS design is managing the large feature space encompassing both product and customer attributes, coupled with the dynamic nature of real-world markets. The continuous introduction of new products, changing customer preferences, and evolving business environments create constant model instability. A related problem is interest drift, referring to the gradual shift in customer interests and purchasing behavior, which makes it difficult for static models to maintain accurate personalization.

Furthermore, the absence of explicit rewards for certain actions limits the ability to establish direct causal relationships between recommendations and actual purchases. Evaluating RS solely based on accuracy metrics is therefore insufficient; models must also be assessed using business-oriented indicators such as profitability, customer retention, and long-term value.

In addition, RS must effectively handle the “long-tail” phenomenon by recommending both popular and niche products. This balance ensures diversity and prevents over-recommendation of high-frequency items, which can reduce user engagement over time.

2.4.3. Limitations of Supervised Methods

Traditional supervised learning methods for RS suffer from several structural limitations. The first is system bias, as models are trained only on observed user interactions with previously recommended items, while potential preferences for unseen items remain unmeasured. This leads to an incomplete understanding of user behavior and reduces generalization capability.

Moreover, supervised approaches generally focus on short-term metrics, overlooking users’ long-term engagement and satisfaction. They also lack mechanisms for real-time updates, continuous learning, and adaptive feedback. In contrast, RL overcomes these constraints by continuously updating its policy through ongoing interaction with the environment, optimizing long-term objectives rather than immediate outcomes.

2.4.4. RL Approaches

RL introduces several advantages for RS, including the integration of user preferences into the objective function, long-term optimization, continuous learning, and lower cumulative regret. However, implementing RL in RS presents several challenges. These include designing appropriate state representations, formulating a valid Markov framework, handling large and complex action spaces (where thousands of items may need to be ranked), managing high exploration costs, and dealing with unobserved user states since preferences are rarely expressed explicitly. Furthermore, RL models often suffer from noisy or sparse reward signals [16].

Despite the rapid development of DRL, several technical obstacles remain. These include the need for large training datasets, the difficulty of designing robust reward functions, sparsity of reward signals, hardware limitations, and dependencies on external factors and hyperparameter configurations [16].

Nevertheless, recent DRL advancements have significantly improved RS performance by enabling dynamic adaptation to user feedback and preferences. DRL frameworks—particularly those based on the MDP—have shown strong potential for managing complex and evolving user–item interactions [17]. By combining these methods with DP strategies, businesses can design RS that not only adapt to user behavior but also align with customers' WTP. Leveraging DRL thus allows firms to optimize pricing and promotional decisions in real time, directly enhancing both sales performance and revenue growth [17].

2.4.5. Incorporating Price in RS

The most significant gap in the field of RS, which renders these systems sub-optimal, is the failure to consider price in their outputs. The price of a recommended product greatly influences whether customers decide to purchase it or not. In this research, the aim is to recommend the appropriate product along with a price aligned with customers' WTP, thereby optimizing the effectiveness of the RS.

In this context, modeling and evaluating the problem where the output includes both the purchase decision and the price, as well as devising a solution method based on the problem's characteristics, are additional challenges.

In recent years, several advanced approaches have been proposed for enhancing recommender systems. Wang et al. provide a comprehensive survey on reinforcement learning for recommendation (RL4RS), highlighting the role of state/action/reward design and the importance of tackling long-term challenges such as delayed rewards, exploration–exploitation, and off-policy evaluation [74]. Alongside RL-based methods, graph contrastive learning has also gained attention. For instance, XSimGCL demonstrates that simple noise-based perturbations of user–item embeddings can provide effective contrastive signals on user–item graphs, yielding consistent improvements in ranking quality across benchmarks [75]. More recently, generative methods such as DiffRec frame recommendation as a denoising diffusion process, enabling the modeling of complex preference distributions while improving scalability and temporal learning [76]. While these approaches have advanced the field in different directions, they primarily focus on ranking accuracy or preference modeling rather than directly incorporating pricing decisions. Our framework introduces a novel 5-class output that jointly recommends products and personalized prices, for which no directly comparable baseline exists in prior work. To enable a fair comparison with existing studies—such as RL4RS, XSimGCL, and DiffRec—that evaluate binary purchase versus non-purchase classification, we additionally map our 5-class predictions into a 2-class setting and report results against these baselines.

3. Proposed Method

This section presents the proposed model for product recommendation aligned with each customer's WTP, developed within an RL framework.

Each customer is represented by a history of interactions that include purchases across different product categories and price ranges, both with and without discounts, as well as expressed preferences and product reviews. In addition to behavioral data, demographic attributes such as age, gender, education level, and income provide essential context for personalization. Correspondingly, each product is characterized by attributes including price range, category, subcategory, discount rate, and sales volume.

The primary objective of the proposed model is to enhance product sales and increase business profitability through personalized recommendations, thereby strengthening customer loyalty. From a marketing perspective, designing effective promotional strategies requires delivering offers that maximize both customer satisfaction and retention. The proposed framework addresses precisely this goal by learning optimal recommendation and pricing strategies tailored to individual purchasing behaviors.

Specifically, the model seeks to recommend the most appropriate product at a price consistent with the customer's WTP. It achieves this by leveraging the joint information contained in customer and product attributes as well as their interaction histories. The system is designed to maximize the likelihood of purchase, thereby improving both organizational profitability and customer experience.

In line with the concept of value-aware recommender systems (VARs), which explicitly integrate business value and user utility into the recommendation process [13], the proposed model incorporates DP strategies within an RL framework. This integration enables the system to simultaneously optimize for customer satisfaction and business profitability in real time.

3.1. Problem Definition

This paper presents a model of product recommendation along with pricing tailored to customer purchase enthusiasm using an RL approach.

The problem specifications for the proposed model include several key features. Customer features encompass purchase history, tracking the variety of purchases, frequency of discounts used, and preferred categories, along with behavioral metrics like discount rates, purchase frequency, and time intervals between discounted purchases. Product features focus on price rank within its category, sales metrics, including the ratio of discounted sales to total sales, and customer diversity in purchases. WTP insights involve analyzing customer data to infer price sensitivity and purchasing power, and understanding product positioning to align with different customer segments.

Unlike traditional RSs that primarily focus on suggesting products based on user preferences, our system integrates pricing strategies. It considers the customer's WTP, ensuring that recommendations are not only relevant but also economically appealing. This approach addresses both the demand and pricing sensitivity, providing a more comprehensive solution.

Our system can be applied in various marketing strategies, such as:

- **Personalized Promotions:** Offering discounts and promotions tailored to individual customers, increasing the likelihood of purchase.
- **DP:** Adjusting prices in real-time based on customer behavior and market trends.
- **Customer Retention Programs:** Creating loyalty programs that offer personalized discounts and exclusive deals, enhancing customer retention.

Price is a critical factor in marketing as it directly influences purchase decisions. For instance, a customer who frequently buys products at discounted prices is less likely to

purchase items at full price. By analyzing this behavior, businesses can offer tailored discounts to encourage purchases, thus increasing sales volume and customer loyalty [12].

Given customer and product features, it might be inferred that a customer who buys 90% of their products at a discount is very unlikely to purchase a full-priced product. Similarly, products priced at the higher end of the price range may not appeal to customers who usually shop in a lower price range. Another example, as depicted in Figure 6, focuses on price-related features: for a customer characterized by a 35% discount rate and a preferred price range of 50 to 70, offering a product priced at 100 with a 40% discount might have a higher likelihood of a successful purchase, and this customer is likely to engage more frequently in the future.



Figure 6. Product Offer Aligned with Customer's Preferred Price Range.

Additionally, if a product has typically sold at a 30% discount, then offering this product at full price would likely have a lower chance of success (Figure 7).



Figure 7. Customized Discount Rate Offer Aligned with Customer's WTP.

In addition to the problem definition, it is important to highlight the generalization capability of the proposed framework. The reinforcement learning approach employed in this study is domain-agnostic, as it is based on fundamental state–action–reward dynamics rather than assumptions specific to the retail sector. With appropriate retraining on sector-specific features and reward functions, the model can be adapted to other industries such as fashion, electronics, or digital services, where consumer behavior may vary in terms of purchase frequency, price sensitivity, and seasonal effects. Prior research has shown that RL-based recommendation systems can successfully generalize across different domains when exposed to new interaction data [15,77].

Furthermore, even within a single sector, product categories display diverse behavioral patterns. During the feature extraction process, a wide range of features was initially considered—derived from both business knowledge and the academic literature—including category-related features such as the number of purchases across different product types.

However, statistical analysis and importance testing revealed that these category-level features exhibited relatively lower discriminative power compared to more fundamental behavioral variables such as purchase frequency, discount usage, and average spending. Therefore, the final feature set prioritized features with higher predictive strength, enabling a more robust modeling of customer behavior and willingness to pay. Nonetheless, the model remains flexible and can be retrained with category-specific features to capture finer distinctions when required.

We propose an RL-based approach to develop a dynamic recommender system that adjusts product prices in real-time according to each customer's WTP. The system uses historical interaction data to learn optimal pricing strategies that maximize long-term customer engagement and revenue.

3.2. Problem Modeling in the RL Framework

In the proposed model, the problem of simultaneous product recommendation and pricing is addressed through an RL approach. In this setting, the optimization process is conducted within the RL paradigm, enabling the model to recommend both the most suitable product and its corresponding price in a unified manner. Owing to the characteristics of RL algorithms—such as their ability to incorporate user preferences into the objective function, maintain a long-term perspective, accumulate rewards over time, and continuously update through learning—the proposed method is expected to achieve superior performance compared to traditional techniques.

The overall schematic of the model, from input to output, is summarized in Figure 8 and described below:

1. **Inputs:** Customer and product features are extracted from the dataset, capturing behavioral interactions between customers and products.
2. **Processing:** Based on customer, product, and interaction data, multiple behavioral and transactional features are computed. The most influential variables are then identified and selected for modeling.
3. **RL Framework:** Using the extracted features, the RL environment is formulated in terms of *states*, *actions*, and *rewards*. The agent is trained using the DQN algorithm to learn the optimal pricing strategy.
4. **Outputs:** The trained agent generates personalized product and price recommendations, including optimal discount levels for each customer.

To model the problem within the RL framework, the MDP formulation is adopted. This involves defining the set of states, actions, transition probabilities, rewards, and the discount factor. In this study, the state space represents the joint information of the customer and product, including their respective attributes and price levels. In other words, a state reflects whether a specific customer, characterized by certain features, purchases a particular product at a given price.

Within this framework, the state captures the combination of customer and product attributes, while actions correspond to recommending specific product–price pairs. The reward function evaluates the success of each recommendation based on purchase likelihood and customer satisfaction, guiding the agent to maximize long-term profitability and customer loyalty.

The RL approach allows the model to adapt and improve continuously by accounting for both immediate and long-term effects of pricing and product recommendations. This dynamic capability effectively addresses the limitations of traditional supervised learning methods, which typically optimize short-term outcomes.

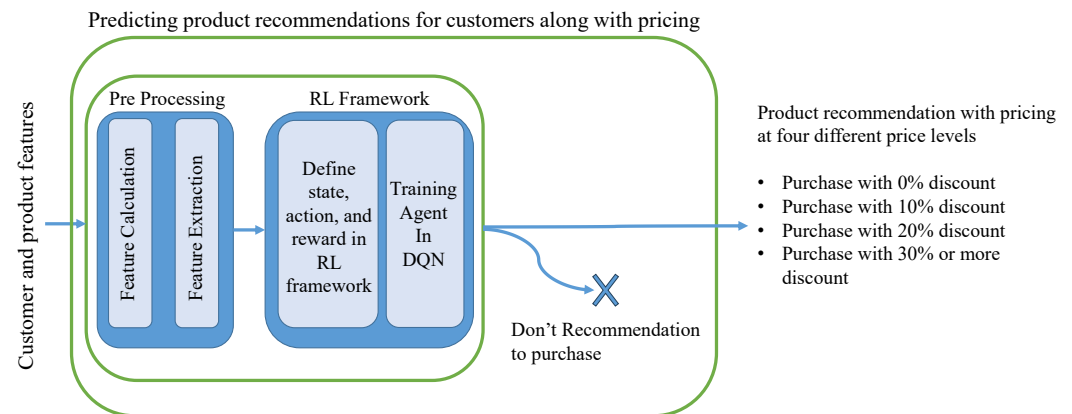


Figure 8. Product Recommendation and Pricing Model with RL Approach.

To better understand the influence of price on customer behavior and improve the interpretability of the system, the first step involves creating comprehensive profiles for both customers and products. These profiles facilitate exploratory data analysis, providing deeper insights into purchasing behavior and serving as the foundation for subsequent modeling stages.

Using customer, product, and interaction data, a diverse set of behavioral and transactional features is computed to describe relationships between customers and products. To identify the most informative variables, clustering is first applied to customer and product features, followed by feature importance analysis using decision trees. The resulting subset of high-impact variables constitutes the state space of the RL environment.

The list of features included in the state space is shown in Table 1.

To mitigate the uncertainty inherent in implicit interaction data, the feature engineering and selection process was designed to enhance the model’s ability to infer customer preferences and WTP with higher precision. Beyond individual purchase or viewing actions, the selected features capture broader behavioral patterns such as discount sensitivity, spending consistency, and purchase frequency across product categories. By integrating these aggregated and statistically validated indicators into the RL state space, the framework minimizes the influence of noise or ambiguity arising from isolated events. Consequently, the RL agent learns from stable and behaviorally meaningful representations of customer activity rather than from raw, potentially misleading signals.

Scalability in the proposed framework is achieved through efficient state and action design. The model defines state and action representations based on aggregated and feature-engineered behavioral attributes instead of product-specific embeddings, allowing it to remain computationally efficient as the catalog size increases. Moreover, using a discrete action space with four discount levels and a no-purchase option reduces the computational burden compared to continuous or product-specific pricing, enabling faster convergence and stable training performance.

The action space includes five states: no purchase, purchase with 0% discount, purchase with 10% discount, purchase with 20% discount, and purchase with more than 30% discount. Table 2 illustrates the action space.

The reward function is defined as follows: if the agent correctly predicts a purchase or no purchase, it receives a reward of 4 units. If a purchase is made and the discount percentage is also correctly predicted, an additional reward of 6 units is given. Otherwise, the agent receives a penalty of −10 units. The reward function is summarized in Table 3.

The reward function in our framework was designed heuristically but with explicit grounding in business logic. The numerical reward and penalty values were selected through an iterative process, tested across several configurations to identify those that

maximized the simulated business profit. The values were assigned asymmetrically to reflect real-world business costs: false negatives (missed purchases) received stronger penalties, correct purchases were rewarded more heavily, and intermediate outcomes were scaled proportionally. This design ensures that the agent’s learning incentives align with long-term profit maximization.

Table 1. List of features forming the state space.

Product Features	
S1	Original product price
S2	Number of unique product sales
S3	Average original product price
S4	Number of unique customers who bought the product
S5	Average time interval between product sales
Customer Features	
S6	Total number of purchase invoices
S7	Customer lifetime
S8	Total number of unique products purchased
S9	Average amount paid
S10	Average actual amount of purchased products
S11	Average discount amount used
S12	Average time interval between all purchases
S13	Average time interval between discounted purchases
S14	RFM score
Final Decision Feature	
S15	Discount rate of the product purchased by the customer

Table 2. Action Space.

Action Code	Description
A1	No purchase
A2	Purchase with 0% discount
A3	Purchase with 10% discount
A4	Purchase with 20% discount
A5	Purchase with more than 30% discount

The selection of the final reward structure was conducted through a comprehensive sensitivity analysis and profit-based evaluation, in which multiple reward tables—each representing different trade-offs between sales volume and profit margin—were tested, and the configuration that achieved the highest cumulative profit and most stable learning convergence was identified as optimal.

Table 3. Reward Function Corresponding to Agent Actions.

Reality	Agent Action	Reward
No purchase	No purchase	4
Purchase with $n\%$ discount	Purchase with $m\%$ discount ($n \neq m$)	4
Purchase with $n\%$ discount	Purchase with $n\%$ discount	10
Purchase with $n\%$ discount	No purchase	−10
No purchase	Purchase with $n\%$ discount	−10

This reward structure incentivizes the agent to accurately predict both the occurrence of a purchase and the correct discount rate, thus optimizing the RS's effectiveness.

3.3. Training the Model

This model uses an RL approach to simultaneously recommend products and their prices. The framework described includes defining states, actions, transition probabilities, and rewards associated with specific actions in specific states. The dataset used for training and evaluating the model comprises 1,946,799 samples, including both positive and negative interactions. Positive samples indicate successful transactions where the recommended product was purchased at the suggested price, while negative samples represent unsuccessful recommendations. This large dataset ensures that the model can learn diverse patterns and generalize well to new data. The neural network architecture for the RL agent is based on a DQN. The DQN model utilizes deep learning to approximate the Q-value function, which represents the expected cumulative reward of taking a given action in a particular state.

- **Input Layer:** The input layer receives the state representation, which includes customer and product features. These features are encoded into a fixed-size vector to be processed by the neural network.
- **Hidden Layers:** Four hidden layers with ReLU activation functions process the input data. ReLU is chosen for its ability to introduce non-linearity while maintaining computational efficiency, preventing vanishing gradient problems common in deep networks.
- **Output Layer:** The output layer produces the Q-values for each possible action, representing the expected rewards for each action given the current state. This layer provides the basis for the decision-making process in the RL framework.

The detailed training procedure for the DQN-based DP model is presented in Algorithm 2.

Algorithm 2 DQN Training for DP

```

1: Initialize online network  $Q_\theta$ , target network  $Q_{\bar{\theta}} \leftarrow Q_\theta$ , replay buffer  $\mathcal{D}$ , discount  $\gamma$ ,
   exploration rate  $\epsilon$ , minibatch size  $B$ , target update period  $C$ .
2: for each episode do
3:   Initialize state  $s_0$ .
4:   for  $t = 0, 1, \dots$  do
5:     Action ( $\epsilon$ -greedy): with prob.  $\epsilon$  choose random  $a_t$ , else  $a_t = \arg \max_a Q_\theta(s_t, a)$ .
6:     Environment step: execute  $a_t$  (price), observe reward  $r_t$ , next state  $s_{t+1}$ , and
       terminal flag  $done_t$ .
7:     Store transition: push  $(s_t, a_t, r_t, s_{t+1}, done_t)$  into  $\mathcal{D}$ .
8:     Minibatch update: sample  $\{(s_i, a_i, r_i, s'_i, done_i)\}_{i=1}^B$  from  $\mathcal{D}$ .
9:     Targets: for each  $i$  set  $y_i = r_i + \gamma(1 - done_i) \max_{a'} Q_{\bar{\theta}}(s'_i, a')$ .
10:    Optimize: update  $\theta$  by minimizing  $\frac{1}{B} \sum_{i=1}^B \ell(y_i - Q_\theta(s_i, a_i))$ 
11:    if  $t \bmod C = 0$  then
12:      Target update:  $\bar{\theta} \leftarrow \theta$ .
13:    end if
14:    Exploration schedule: decay  $\epsilon$ .
15:    if  $done_t$  then
16:      break
17:    end if
18:  end for
19: end for

```

The final DQN configuration was determined based on established guidelines and empirical testing. A learning rate of 1×10^{-3} (Adam optimizer), discount factor $\gamma = 0.95$,

replay buffer of 50,000 transitions, mini-batch size of 64, and target network updates every 1000 steps ensured stable convergence. The MSE loss was selected over alternative formulations such as the Huber loss, as it provided smoother convergence in our DP environment with relatively stable reward distributions. This setup yielded consistent results across multiple runs.

The trained model is evaluated using the test dataset. The results show significant improvements in recommendation accuracy and pricing strategy compared to traditional methods. The RL approach effectively adapts to customer behavior, providing personalized recommendations and optimal pricing strategies. One of the key advantages of using RL in this context is the model's ability to continuously learn and adapt. As new data becomes available, the model can be retrained to incorporate recent trends and changes in customer behavior, ensuring that the recommendations and pricing strategies remain relevant and effective. Continuous learning allows the model to stay up-to-date with the latest customer preferences and market conditions, providing a dynamic and responsive RS.

3.4. Evaluation Criteria

Evaluation metrics are defined and selected according to the problem structure and its outputs. In this problem, product recommendations along with pricing are provided. The output of the problem is defined as multi-class. Therefore, multi-class classification metrics such as macro average accuracy, macro average recall, macro F-score, micro F-score [78], One-vs-Rest (OvR) macro averaged AUC [79] and Normalized Discounted Cumulative Gain (NDCG) are used. Additionally, weighted multi-class precision is proposed for evaluation based on different preferences.

Considering that the model output is a type of multi-class classification output, multi-class evaluation methods are used to evaluate this model. The confusion matrix for multi-class classification for one sample class is shown in Table 4, and the metrics of macro average accuracy, macro average recall, macro F-score, micro F-score, macro averaged AUC and NDCG are used to evaluate this model.

Table 4. Metrics in Multi-Class Classification for the Class ‘Purchase with 10% Discount’.

Actual	Predicted				
	No Purchase	0% Discount	10% Discount	20% Discount	≥30% Discount
No Purchase	TN	TN	FP	TN	TN
0% Discount	TN	TN	FP	TN	TN
10% Discount	FN	FN	TP	FN	FN
20% Discount	TN	TN	FP	TN	TN
≥30% Discount	TN	TN	FP	TN	TN

A new evaluation metric is proposed such that, based on the weights given in Table 5, each sample that falls into the specified condition is assigned this coefficient, and then the model is evaluated accordingly.

The coefficients in Table 6 emphasize that if the correct class predictions of various non-purchase or purchases with different discounts are made, a coefficient of +10 is considered. In the case of any purchase with a predicted non-purchase, the highest negative coefficient of −4 is considered. If the purchase is made at different prices and the purchase is predicted at a price different from the actual purchase category, a negative coefficient is applied accordingly. The closer the predicted discount category is to the actual purchase category, the smaller the negative coefficient. For example, if the purchase is from the 20% discount category, a non-purchase prediction gets the highest negative coefficient, and the correct

20% discount category prediction gets a coefficient of +10. Predicting a higher discount category has a smaller negative coefficient because offering a higher discount might slightly reduce business profit, but predicting a lower discount category has a higher likelihood of not leading to a purchase and thus gets a higher negative coefficient.

Table 5. Weights for Predicting Different Classes to Calculate Weighted Multi-Class Accuracy.

Actual	Prediction				
	No Purchase	0% Discount	10% Discount	20% Discount	≥30% Discount
No Purchase	10	−4	−3	−2	−1
0% Discount	−4	10	−1	−2	−3
10% Discount	−4	−3	10	−1	−2
20% Discount	−4	−3	−2	10	−1
≥30% Discount	−4	−3	−2	−1	10

Table 6. Dunnhumby dataset specifications.

Feature	Value
Number of Transactions	2,595,732
Number of Customers	2500
Number of Products	92,339
Time Period	711 days
Number of Invoices	276,484
Number of Product Categories	308
Number of Product Subcategories	2383

When the number of samples in each of the above conditions is obtained and multiplied by the coefficients specified in Table 6, the ‘weighted multi-class accuracy’ metric is defined and calculated as follows

$$\text{Weighted Multi-Class Accuracy} = \frac{\text{Sum of all positive and negative values}}{\text{Total number of samples} \times 10} \quad (1)$$

Due to the absence of a reference for comparing 5-class results in studies, the product recommendation results along with pricing were converted to binary outputs: purchase and non-purchase. These binary results were then compared with the reference studies.

3.5. Challenges and Innovations

The key challenges encountered during the development of the proposed model, along with its main innovations, are outlined below.

- **Challenges**

1. **Defining the Problem by Simultaneously Considering Price and Product Recommendation:** A major challenge involved formulating the problem in a way that simultaneously incorporates both product recommendation and pricing. Traditional RSs typically treat these two elements independently; however, our framework required their joint optimization within a single model. This integration ensures that recommendations are not only relevant but also appropriately priced to align with customers’ WTP and purchasing intent.
2. **Selecting an Appropriate Solution Based on Problem Characteristics:** Choosing the most suitable methodological approach was critical due to the multifaceted nature of the problem. The selected solution needed to effectively

manage the complexities of personalized pricing, heterogeneous customer preferences, diverse product attributes, and DP strategies, while maintaining scalability and interpretability.

3. **Challenges of Implementing a RL Approach:** Developing an RL-based system introduced additional technical challenges, including the design of an appropriate state space, action space, and reward function tailored to the specific requirements of the problem. Moreover, the iterative learning process inherent in RL requires large volumes of high-quality training data, computational resources, and careful tuning of hyperparameters to ensure convergence and stability.

- **Innovations**

1. **Novel and Business-Relevant Problem Definition:** The proposed research addresses a novel and practically significant problem by integrating pricing into the recommendation process. This addition directly enhances the ability of RS to influence purchasing decisions, thus overcoming a key limitation of conventional RS.
2. **High Business Applicability and Marketing Optimization:** The approach provides a practical and scalable solution for businesses by optimizing marketing expenditures and improving promotional effectiveness. Through tailored product and price recommendations, companies can enhance customer loyalty, increase retention rates, and generate sustainable long-term value.
3. **Financial Evaluation and Direct Economic Impact:** The model establishes a direct link between its outputs and financial metrics, enabling businesses to evaluate the monetary impact of their recommendation strategies. This connection facilitates evidence-based decision-making and allows for precise assessment of profitability gains resulting from the system's deployment.
4. **Problem Formulation and Modeling Aligned with Domain Characteristics:** The problem was carefully formulated and modeled to reflect its unique structural and behavioral characteristics. The proposed framework was explicitly designed to handle the complexity of personalized pricing and recommendation, ensuring realistic and data-driven optimization.
5. **Integration of Reinforcement Learning within an MDP Framework:** By framing the problem as a MDP and solving it using RL techniques, the model achieves adaptive, feedback-driven learning based on customer interactions. This enables continuous improvement in recommendation quality and pricing precision.
6. **Development of Problem-Specific Evaluation Metrics:** New evaluation metrics were designed to capture both recommendation accuracy and pricing effectiveness. These customized indicators allow a comprehensive assessment of the model's performance in terms of predictive precision and economic relevance.
7. **Adaptability to Business Objectives through Reward Function Adjustments:** The model demonstrates high adaptability to changing business strategies by allowing modifications to the reward function. For instance, the system can be configured to discourage low-margin recommendations, prioritize high-value purchases, or balance discount levels according to marketing objectives. This flexibility ensures alignment with evolving business priorities and operational contexts.

By addressing these challenges and introducing the above innovations, the proposed RL-based model provides a comprehensive and adaptable solution for product recommendation and pricing aligned with customers' WTP. The framework not only meets current business needs but also offers scalability and flexibility to accommodate future market changes and strategic developments.

4. Experiments

This section provides a comprehensive analysis of the experimental results, starting with a detailed description of the dataset used for training and evaluation. The section then outlines the evaluation methodology used to assess the model's effectiveness. The methodology includes training the DQN model, validating its performance, and testing it on a separate dataset. Performance metrics such as precision, recall, and F1 score are used to evaluate the model's accuracy and relevance of recommendations. Finally, we analyze the experimental results, demonstrating significant improvements in recommendation accuracy and pricing strategy compared to traditional methods, and highlight the model's ability to adapt to customer behavior through continuous learning.

4.1. Dataset Description

The Dunnhumby dataset is selected due to its relevance to the problem at hand and its extensive use in numerous studies. This dataset adequately covers features related to discounts, pricing, and promotional attributes. The Dunnhumby dataset is an extensive collection of retail transaction data, encompassing 2,595,732 purchase transactions from 2500 unique customers over 711 days. It includes detailed customer demographics, product attributes, and transactional specifics such as pricing and discounts. This dataset is instrumental for developing and testing machine learning algorithms, particularly in customer segmentation, product recommendation, DP, and market basket analysis. Its comprehensive nature allows for robust real-world applications, enhancing retail strategies and customer satisfaction through data-driven insights (<https://www.dunnhumby.com/>, accessed on 9 September 2025).

Some specifications of the Dunnhumby dataset are presented in Table 6.

To the 2,595,732 positive samples in the dataset, 3,893,598 negative samples (or non-purchase instances) are added, making the total dataset size 6,489,330 samples. Thus, the ratio of positive samples to the total is 2 to 5. In other words, 60% of the samples are non-purchase instances, and 40% are purchase instances with various discount percentages. Considering 30% of the samples as test samples, the number of test samples will be 1,946,799, the details of which are shown in Table 7.

Table 7. Train and test split.

Category	Percentage	Number of Transactions
Training	70%	4,542,531
Testing	30%	1,946,799

4.2. Experimental Results

The evaluation of this model utilizes multi-class classification metrics and the proposed weighted multi-class accuracy based on Table 5. Figure 9 shows the confusion matrix for this modeling approach.

The confusion matrix presented in our study showcases the performance of an RL model designed for product recommendation and pricing tailored to customer purchase enthusiasm. The true labels on the Y-axis represent the actual customer behaviors, while the predicted labels on the X-axis indicate the model's predictions. Each cell's value represents the count of instances for each combination of true and predicted labels.

Our model demonstrates a high degree of accuracy in predicting non-purchasing behavior, with a substantial true positive count of 1,063,240 in the "Don't Purchase" category. This high accuracy is critical as it allows the model to effectively identify customers who are unlikely to make a purchase, thereby enabling the business to minimize marketing expenditures on these non-responsive segments.

The ability to predict high discount categories accurately is a crucial strength of the model. This accuracy ensures that the model can effectively identify customers who are most likely to respond only when offered significant incentives. In the context of marketing and sales, different customers exhibit varying levels of price sensitivity. Some customers might make a purchase without any discount, while others might require substantial discounts to be persuaded to buy.

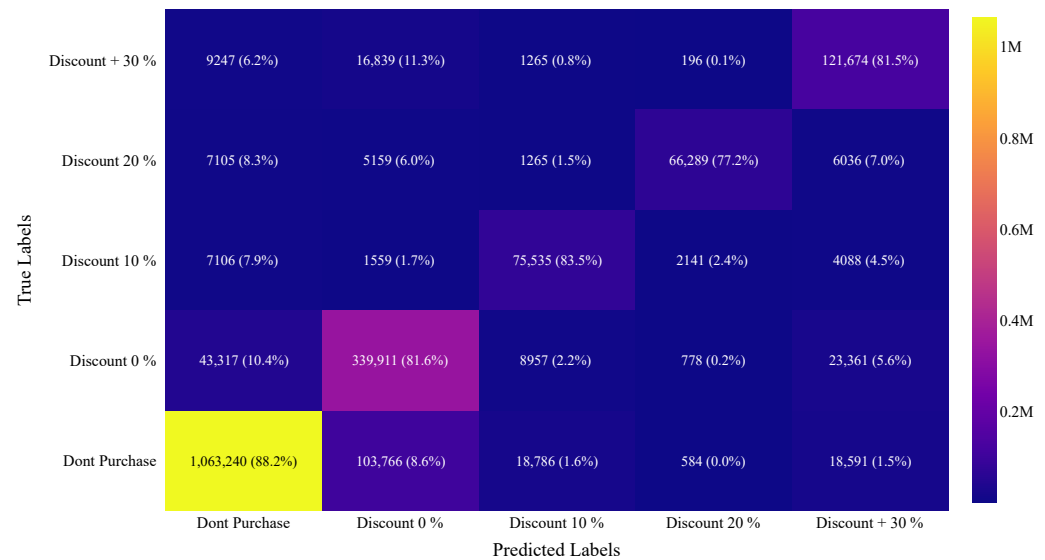


Figure 9. Confusion Matrix for the RL-Based Pricing Recommendation Model.

When the model accurately predicts that a customer falls into a high discount category (such as 20% or 30%), it indicates that the customer has a high price sensitivity. These customers are less likely to convert without significant incentives. By correctly identifying these customers, the business can tailor its marketing efforts more effectively. Offering substantial discounts to these customers can lead to successful conversions that might not have occurred otherwise.

Moreover, precise identification of high discount customers helps in optimizing the allocation of marketing resources. Instead of blanket discount offers to all customers, the business can target high-discount incentives specifically to those who need them. This targeted approach not only improves conversion rates but also enhances overall profitability. It ensures that the business does not erode its margins by offering unnecessary discounts to customers who would have purchased at a lower discount or even at full price.

Additionally, accurate prediction of high discount categories can improve customer satisfaction. Customers who receive personalized offers that match their price sensitivity are more likely to feel valued and understood, enhancing their overall experience with the brand. This positive experience can lead to increased customer loyalty and long-term customer relationships.

In summary, the model's ability to accurately predict high discount categories is vital for identifying customers who need significant incentives to convert. It allows for targeted marketing strategies, optimized resource allocation, improved conversion rates, enhanced profitability, and better customer satisfaction. This precision in prediction supports the overall effectiveness of the personalized pricing strategy, making it a valuable asset for the business.

The evaluation results for the proposed RL-based pricing and product recommendation model are presented in Table 8.

Table 8. Evaluation Results for the RL-Based Pricing Recommendation Model.

Metric	Test Data
Macro Average Precision	0.8059
Macro Average Recall	0.8243
Macro F-score	0.8108
Micro F-score	0.8560
Macro Averaged AUC	0.8743
NDCG@5	0.8947
Weighted Multi-class Accuracy	0.8082

To comprehensively assess the performance of the proposed RL model for product recommendation and pricing, several key metrics were analyzed on the test data, including Macro Average Precision, Macro Average Recall, Macro and Micro F-scores, Macro Averaged AUC, NDCG@5, and Weighted Multi-class Accuracy. Together, these indicators provide a holistic view of the model's capability to handle multi-class classification and ranking tasks.

A **Macro Average Precision** of 0.8059 indicates that, on average, 80.59% of the predicted instances were correctly classified. This high precision demonstrates that the model effectively minimizes false positives across all discount classes, ensuring that both recommendations and price adjustments are accurately tailored to the appropriate customer segments.

The **Macro Average Recall** of 0.8243 shows that, on average, 82.43% of the actual instances for each class were correctly identified. This strong recall value suggests that the model successfully captures most relevant cases across different classes, minimizing the number of customers whose purchasing intentions are overlooked by the RS.

The **Macro F-score** of 0.8108, representing the harmonic mean of Macro Precision and Recall, reflects the model's balanced and consistent performance across all discount categories. This equilibrium between precision and recall is crucial in RS, where both the accuracy of suggested products and the inclusiveness of relevant options are important.

The **Micro F-score** of 0.8560 aggregates performance across all classes and accounts for class imbalance, reflecting the model's overall accuracy at the dataset level. This result confirms that the model performs robustly across diverse customer behaviors and varying discount preferences.

The **Macro Averaged AUC** of 0.8743 further validates the strong discriminative ability of the proposed model across the five discount classes. Because AUC is a threshold-independent measure, this result indicates that the model consistently distinguishes between high and low purchase probabilities, regardless of classification boundaries. The macro-averaged computation ensures that all classes contribute equally, confirming a reliable ability to rank customers' purchase likelihoods across multiple discount levels.

The **NDCG@5** score of 0.8947 highlights the model's exceptional ranking performance across the ordered discount categories. This metric measures how well the model prioritizes desirable outcomes—such as purchases at lower discount rates—while penalizing suboptimal ones, such as unnecessary deep discounts or missed opportunities. The high NDCG@5 score demonstrates that the model ranks pricing actions in a manner consistent with business profitability goals.

Finally, the **Weighted Multi-class Accuracy** of 0.8082 reflects the model's predictive effectiveness when accounting for predefined class importance. This confirms that the system achieves reliable performance even when different classes (e.g., discount levels) carry different business priorities.

Collectively, these results demonstrate that the proposed RL model performs at a high level of accuracy, consistency, and business relevance. The strong Macro and Micro F-scores confirm balanced predictive capability across classes, while the high AUC and NDCG@5 values underline the model's superior ranking and discriminative power. These findings validate the model's effectiveness in jointly optimizing product recommendations and pricing decisions in alignment with customer purchasing behavior. Consequently, the proposed RL-based framework represents a valuable and practical tool for enhancing marketing decision-making, increasing profitability, and improving customer satisfaction.

The detailed Precision, recall, and F-score macro for the five output classes are presented in Tables 9, 10, and 11, respectively.

Table 9. Precision Evaluation for Classes in the RL-Based Pricing Recommendation Model.

Classes	Precision
Class 1—No Purchase	0.9409
Class 2—Purchase at Full Price	0.7270
Class 3—Purchase with 10% Discount	0.7130
Class 4—Purchase with 20% Discount	0.9471
Class 5—Purchase with 30% Discount or More	0.7002

The model exhibits high precision for predicting non-purchase behavior (94.09%) and purchases with a 20% discount (94.71%), indicating that it accurately identifies customers in these categories. This precision is critical for minimizing false positives and ensuring that marketing resources are effectively utilized. However, the precision for predicting full-price purchases (72.7%), purchases with a 10% discount (71.3%), and high-discount purchases (70.02%) is moderate, suggesting areas for improvement to enhance the accuracy of these predictions.

Table 10. Recall Evaluation for Classes in the RL-Based Pricing Recommendation Model.

Classes	Recall
Class 1—No Purchase	0.8820
Class 2—Purchase at Full Price	0.8160
Class 3—Purchase with 10% Discount	0.8352
Class 4—Purchase with 20% Discount	0.7721
Class 5—Purchase with 30% Discount or More	0.8153

In terms of recall, the model demonstrates strong performance in identifying non-purchasers (88.2%), full price purchases (81.6%) and customers who will purchase with a 10% discount (83.52%) or a high discount (81.53%). This indicates that the model effectively captures the majority of relevant instances in these categories. The recall for 20% discount purchases (77.21%) also shows good coverage, though further refinement could help in capturing more instances within these groups.

The Macro F-score, which balances precision and recall, is high for non-purchase (91.07%) and 20% discount (85.07%) categories, reflecting robust overall performance. The F-scores for full-price (76.94%), 10% discount (76.98%), and high-discount purchases (75.34%) indicate a balanced but moderate performance, highlighting potential areas for model enhancement to achieve better accuracy.

In summary, the model performs exceptionally well in predicting non-purchases and 20% discount purchases, with high precision and balanced performance. There are opportunities to improve the precision and recall for full-price, 10% discount, and high-discount categories to optimize resource allocation and marketing efforts. These metrics

validate the model’s capability to effectively tailor product recommendations and pricing strategies while identifying specific areas for further refinement to enhance customer satisfaction and conversion rates.

Table 11. Macro F-score Evaluation for Classes in the RL-Based Pricing Recommendation Model.

Classes	Macro F-Score
Class 1—No Purchase	0.9107
Class 2—Purchase at Full Price	0.7694
Class 3—Purchase with 10% Discount	0.7698
Class 4—Purchase with 20% Discount	0.8507
Class 5—Purchase with 30% Discount or More	0.7534

Due to the absence of authoritative studies that have implemented this proposed method, the results of the modeling method for recommending products along with price and the use of a multi-class output in this discussion, an approach for evaluating this method is to simplify the obtained results into two classes, purchase and non-purchase, and compare them with studies that have used RL-based methods for product recommendation in RS.

Studies [74–76] examined various DRL methods in the application of RS. The best results in a 2-class RS that suggests purchase or non-purchase are those in the experiments reported in these studies. According to the results presented in Table 12, our RL-based approach consistently exceeds the performance of Wang et al, DiffRec and XSimGCL across all key metrics, demonstrating its robustness and effectiveness for two-class purchase prediction.

The confusion matrix of results from the 5-class modeling is shown in Figure 9. In the course of converting the 5-class results into 2-class results, all purchases, whether at full price or with various discounts, are considered as a purchase, and non-purchase is considered as non-purchase, i.e., classes 2 to 5 are merged into each other. With this conversion, the confusion matrix for the 2-class purchase and non-purchase appears as Figure 10.

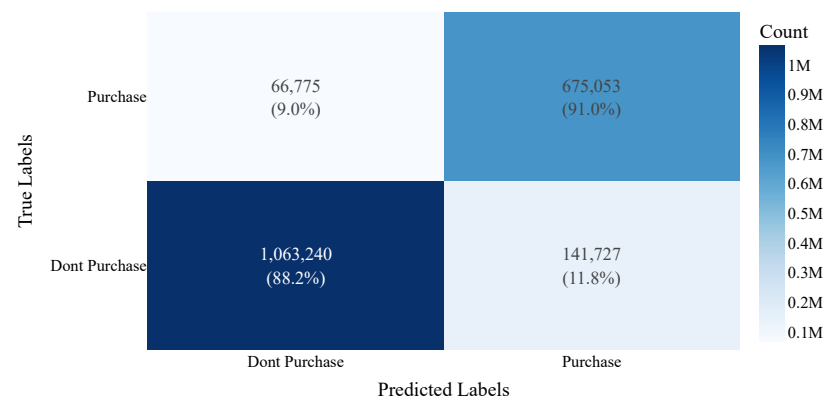


Figure 10. Confusion Matrix for the RL-Based Pricing Recommendation Model—converted to 2-Class.

The results of the evaluation metrics from converting to the 2-class problem of purchase or non-purchase and comparison with the results of studies [74–76] are also shown in Figure 11. These evaluation results show the modeling method for recommending products along with price using an RL approach and a comparison of the converted 2-class results with [74–76].

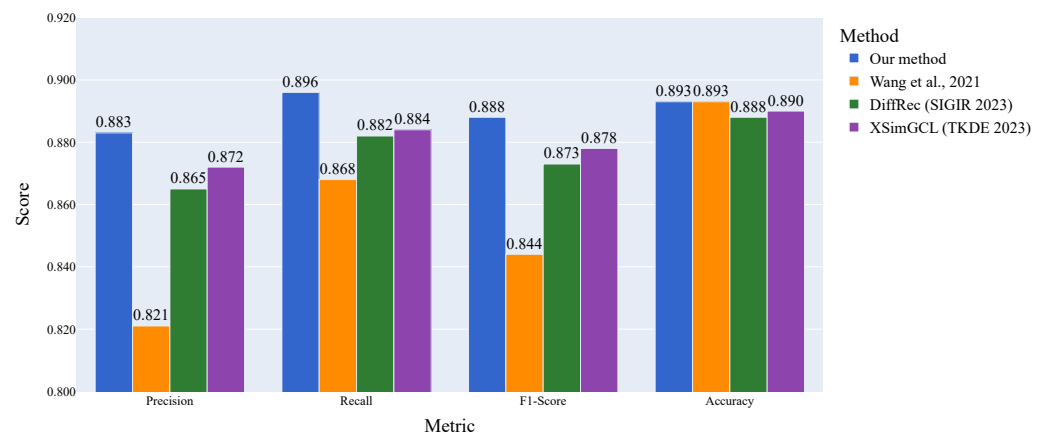


Figure 11. Evaluation results in product recommendation modeling with pricing using a RL approach and comparison of converted to 2-class results with studies [74–76].

Table 12. Comparison with recent state-of-the-art methods. Bold indicates the best score per metric (ties bolded).

Method	Precision	Recall	F1-Score	Accuracy
Our method	0.883	0.896	0.888	0.893
Wang et al., 2021 [74]	0.821	0.868	0.844	0.893
DiffRec (SIGIR 2023) [76]	0.865	0.882	0.873	0.888
XSimGCL (TKDE 2023) [75]	0.872	0.884	0.878	0.890

Our RL-based pricing recommendation model demonstrates superior performance across all key evaluation metrics when compared with three representative state-of-the-art baselines: Wang et al. [74], DiffRec [76], and XSimGCL [75]. In terms of **precision**, our model achieves **0.883**, which is higher than Wang et al.’s **0.821**, DiffRec’s **0.865**, and XSimGCL’s **0.872**, indicating that our recommendations more accurately capture true purchases with fewer false positives. For **recall**, our model attains **0.896**, exceeding the **0.868** of Wang et al., **0.882** of DiffRec, and **0.884** of XSimGCL, highlighting its strength in minimizing missed opportunities for conversions. The balanced effectiveness of our approach is reflected in the **F1-Score** of **0.888**, which surpasses the best competing result of **0.878** (XSimGCL). Finally, our model achieves an **accuracy** of **0.893**, which is on par with Wang et al. (**0.893**) and slightly ahead of DiffRec (**0.888**) and XSimGCL (**0.890**). Overall, these improvements underscore the advantages of explicitly modeling pricing as an action and optimizing a profit-aligned reward under willingness-to-pay constraints—capabilities that are not natively supported by diffusion-based or contrastive-learning recommenders.

4.3. Analysis of Experimental Results

The proposed RL-based model for product recommendation and pricing demonstrates consistently strong performance in both multi-class and binary evaluation settings. In the multi-class configuration, the model accurately predicts non-purchases as well as purchases across different discount levels, confirming that the integration of pricing into the recommendation output is both feasible and effective. When simplifying the five-class output into a binary system distinguishing between “Purchase” and “Don’t Purchase,” the model maintains robust performance across all key metrics, reinforcing its reliability in capturing customer purchasing behavior.

When compared with three representative state-of-the-art baselines—Wang et al. [74], DiffRec [76], and XSimGCL [75]—our model consistently outperforms them. While these

prior methods primarily focus on ranking accuracy, preference modeling, or contrastive representation learning, the proposed framework explicitly integrates pricing decisions into the recommendation process and optimizes a profit-aware reward function aligned with customers' WTP. This dual focus enables the model to capture both purchase likelihood and price sensitivity, resulting in more accurate, profit-oriented, and business-relevant recommendations than the existing baselines.

From a marketing perspective, accurately predicting customer purchasing behavior enables more precise and effective promotional campaigns. By identifying which customers are most responsive to specific discount levels, the model supports finer segmentation of the customer base. This targeted approach reduces marketing inefficiency while improving customer satisfaction by delivering personalized discounts that are more likely to convert. The model's high recall ensures that a greater proportion of the potential customer base is reached, maximizing the overall impact of marketing initiatives. Furthermore, customers who receive tailored offers that align with their price expectations are more likely to perceive the brand as attentive and customer-centric, thereby increasing purchase intent and loyalty.

In terms of revenue optimization, the model's predictive strength under various discount scenarios is critical. By identifying the optimal discount level required to convert each customer segment, the model helps businesses establish pricing strategies that maximize revenue and profitability. Delivering the right discount to the right customer at the right time can substantially increase sales volume while preserving profit margins. For instance, by minimizing unnecessary high discounts and reserving deeper incentives for highly price-sensitive customers, the firm can optimize its overall promotional expenditure. The model's balanced performance, reflected in the F-score of 0.8884 and its strong precision and recall values, confirms its ability to drive revenue growth through improved customer targeting and personalized pricing strategies. This capability not only enhances short-term sales but also fosters long-term customer retention and lifetime value.

Although the model effectively predicts both non-purchasing behavior and high-discount purchase scenarios, some challenges remain in accurately classifying intermediate discount categories. Future improvements could focus on refining the state and action space representations—such as by clustering similar products and customers—to enhance contextual differentiation. Additionally, tuning the reward function to better reflect long-term user engagement and purchase satisfaction could further improve predictive accuracy. Incorporating larger and more diverse training datasets, performing hyperparameter optimization, and extending episode lengths may also enhance model convergence and generalization. These refinements would improve the system's capacity to distinguish between closely related discount levels, resulting in even more precise and effective personalized pricing strategies.

5. Conclusions and Future Works

This research presented an RL-based framework for personalized product recommendation and pricing, demonstrating that incorporating customers' WTP into the recommendation process can substantially improve both customer satisfaction and business profitability. Experimental results confirm this conclusion, showing strong performance across multiple evaluation metrics—including Precision, Recall, F1-score, AUC, and NDCG—thus validating the model's ability to accurately predict purchase behavior and rank offers according to profitability. Collectively, these findings indicate that the proposed framework effectively aligns recommendation and pricing strategies with long-term business objectives.

This study contributes to the RS and marketing literature by bridging traditional product recommendation and DP within a unified RL framework. The integration of behavioral and transactional features, combined with a reward function optimized for profit maximiza-

tion, provides both theoretical rigor and practical value. From a managerial perspective, the model enhances marketing efficiency by enabling adaptive, data-driven promotional strategies that respond dynamically to customer behavior and market conditions.

Future work will focus on expanding the framework through enhanced pre-processing and post-processing techniques. Improved pre-processing methods will ensure cleaner, more structured input data, while advanced post-processing tools will aid in interpreting model outputs and transforming them into actionable business insights. This will bridge the gap between technical outputs and strategic decision-making, enhancing the practical relevance of the model.

A key limitation of the current framework lies in its discrete action space, which includes four fixed discount levels plus a no-purchase option. These thresholds, derived from the Dunhumby dataset, reflect real-world retail practices where stepwise discounts are prevalent. Although this approach promotes stability and interpretability, it does not fully represent the continuous nature of DP. Future extensions could incorporate continuous action spaces using algorithms such as Deep Deterministic Policy Gradient (DDPG) [80] or Soft Actor–Critic (SAC) [81], enabling more granular and realistic price recommendations.

Another promising direction involves refining the design of the reward function. Instead of relying solely on heuristic definitions, future studies could apply systematic reward modeling approaches, such as Inverse Reinforcement Learning (IRL) [82] or Multi-Objective Reinforcement Learning (MORL) [83], to infer or optimize reward structures directly from observed business profit signals. This would allow for more adaptive, profit-sensitive, and objective-driven learning outcomes.

Temporal dynamics also represent a key area for advancement. Considering the timing between purchases and sequences of consumer behavior could significantly enhance predictive accuracy and allow the model to better anticipate future actions. Incorporating temporal dependencies and seasonality effects would enable the framework to adapt more effectively to evolving market conditions and customer life cycles.

Further extensions may explore hybrid or alternative modeling techniques. For instance, integrating transformer-based architectures could be particularly effective given their proven ability to model sequential dependencies. Leveraging transformers may help capture complex, long-range relationships within customer interactions, further improving predictive power and personalization quality.

Ensuring scalability and adaptability remains critical for practical deployment. Future work should test the model in diverse business environments to validate robustness and generalizability. A scalable framework capable of maintaining performance across multiple contexts will enhance its potential for real-world application.

Finally, developing mechanisms for real-time learning and evaluation constitutes an essential step toward operational implementation. Building systems capable of continuous adaptation to incoming data would enable the model to respond rapidly to changing customer behavior and market dynamics. Real-time decision support would empower businesses to execute data-driven strategies that maximize profitability and customer satisfaction in fast-moving retail environments.

In summary, this research establishes a foundation for advanced pricing and RS based on reinforcement learning. Future studies will aim to address current limitations, refine the methodological components, and explore innovative extensions to further improve model performance and applicability. By continuously enhancing the framework with state-of-the-art techniques, this line of research seeks to deliver a powerful, adaptive, and profit-oriented tool capable of supporting sustainable marketing and pricing strategies in dynamic commercial ecosystems.

Author Contributions: Conceptualization, A.M. and B.B.; methodology, A.M. and B.B.; software, A.M.; validation, A.M., B.B. and H.M.; investigation, A.M. and B.B.; writing—original draft preparation, A.M.; writing—review and editing, A.M.; visualization, A.M.; supervision, B.B. and H.M.; project administration, B.B. and H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All of the real graphs used in this paper are publicly available as described in the body of the paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. StartUp Guru Lab. Customer Acquisition vs Retention Costs. 2025. Available online: <https://startupgurulab.com/customer-acquisition-vs-retention-costs> (accessed on 6 August 2025).
2. Optimove. Customer Acquisition vs Retention Costs: Why Retention is More Profitable. 2025. Available online: <https://www.optimove.com/resources/learning-center/customer-acquisition-vs-retention-costs> (accessed on 6 August 2025).
3. Gallo, A. The Value of Keeping the Right Customers. Harvard Business Review. 2014. Available online: <https://hbr.org/2014/10/the-value-of-keeping-the-right-customers> (accessed on 6 August 2025).
4. Chen, K.; Chen, T.; Zheng, G.; Jin, O.; Yao, E.; Yu, Y. Collaborative personalized tweet recommendation. In Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, OR, USA, 12–16 August 2012; pp. 661–670.
5. Kumar, V.; Ashraf, A.R.; Nadeem, W. AI-powered marketing: What, where, and how? *Int. J. Inf. Manag.* **2024**, *77*, 102783. [CrossRef]
6. Kshetri, N.; Dwivedi, Y.K.; Davenport, T.H.; Panteli, N. Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda. *Int. J. Inf. Manag.* **2024**, *75*, 102716. [CrossRef]
7. Nouri-Harzvili, M.; Hosseini-Motlagh, S.M. Dynamic discount pricing in online retail systems: Effects of post-discount dynamic forces. *Expert Syst. Appl.* **2023**, *232*, 120864. [CrossRef]
8. Kim, S.; Park, E. STAD-GCN: Spatial–Temporal Attention-based Dynamic Graph Convolutional Network for retail market price prediction. *Expert Syst. Appl.* **2024**, *255*, 124553. [CrossRef]
9. Lu, C.C.; Wu, I.L.; Hsiao, W.H. Developing customer product loyalty through mobile advertising: Affective and cognitive perspectives. *Int. J. Inf. Manag.* **2019**, *47*, 101–111. [CrossRef]
10. Quadrana, M.; Cremonesi, P.; Jannach, D. Sequence-aware recommender systems. *Acm Comput. Surv.* **2018**, *51*, 1–36. [CrossRef]
11. Zheng, Y.; Wang, D. A survey of recommender systems with multi-objective optimization. *Neurocomputing* **2022**, *474*, 141–153. [CrossRef]
12. Jiang, Y.; Liu, Y. Optimization of online promotion: A profit-maximizing model integrating price discount and product recommendation. *Int. J. Inf. Technol. Decis. Mak.* **2012**, *11*, 961–982. [CrossRef]
13. De Biasio, A.; Montagna, A.; Aiolfi, F.; Navarin, N. A systematic review of value-aware recommender systems. *Expert Syst. Appl.* **2023**, *226*, 120131. [CrossRef]
14. Cantador, I.; Bellogín, A.; Castells, P. In Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems, Chicago, IL, USA, 27 October 2011.
15. Afsar, M.M.; Crump, T.; Far, B. Reinforcement Learning based Recommender Systems: A Survey. *ACM Comput. Surv.* **2022**, *55*, 1–38. [CrossRef]
16. Lin, Y.; Liu, Y.; Lin, F.; Zou, L.; Wu, P.; Zeng, W.; Chen, H.; Miao, C. A Survey on Reinforcement Learning for Recommender Systems. *arXiv* **2021**, arXiv:2109.10665. [CrossRef]
17. Chen, X.; Yao, L.; McAuley, J.; Zhou, G.; Wang, X. Deep reinforcement learning in recommender systems: A survey and new perspectives. *Knowl.-Based Syst.* **2023**, *264*, 110335. [CrossRef]
18. Wen, S.; Shu, Y.; Rad, A.; Wen, Z.; Guo, Z.; Gong, S. A deep residual reinforcement learning algorithm based on Soft Actor-Critic for autonomous navigation. *Expert Syst. Appl.* **2025**, *259*, 125238. [CrossRef]
19. Wang, J.; Karatzoglou, A.; Arapakis, I.; Jose, J.M. Reinforcement Learning-based Recommender Systems with Large Language Models for State Reward and Action Modeling. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), Washington, DC, USA, 14–18 July 2024; pp. 1–11. [CrossRef]
20. Kavoosi, A.; Tavakkoli-Moghaddam, R.; Sajedi, H.; Tajik, N.; Tafakkori, K. Dynamic pricing and inventory control of perishable products by a deep reinforcement learning algorithm. *Expert Syst. Appl.* **2025**, *291*, 128570. [CrossRef]

21. Tanveer, J.; Lee, S.-W.; Rahmani, A.M.; Aurangzeb, K.; Alam, M.; Zare, G.; Malekpour Alamdari, P.; Hosseinzadeh, M. PGA-DRL: Progressive graph attention-based deep reinforcement learning for recommender systems. *Inf. Fusion* **2025**, *121*, 103167. [CrossRef]
22. Sitar-Tăut, D.A.; Mican, D.; Mare, C. Customer behavior in the prior purchase stage – information search versus recommender systems. *Econ. Comput. Econ. Cybern. Stud. Res.* **2020**, *54*, 59–76. [CrossRef]
23. Zhang, H.; Zhao, L.; Gupta, S. The role of online product recommendations on customer decision making and loyalty in social shopping communities. *Int. J. Inf. Manag.* **2018**, *38*, 150–166. [CrossRef]
24. Ramampiaro, H.e.a. New Ideas in Ranking for Personalized Fashion Recommender Systems. In *Business and Consumer Analytics: New Ideas*; Springer: Berlin/Heidelberg, Germany, 2019. [CrossRef]
25. Umberto, P. Developing a price-sensitive recommender system to improve accuracy and business performance of ecommerce applications. *Int. J. Electron. Commer. Stud.* **2015**, *6*, 1–18. [CrossRef]
26. Bai, T.; Wang, X.; Zhang, Z.; Song, W.; Wu, B.; Nie, J.Y. GPR-OPT: A Practical Gaussian optimization criterion for implicit recommender systems. *Inf. Process. Manag.* **2024**, *61*, 103525. [CrossRef]
27. Pujahari, A.; Sisodia, D.S. Modeling users' preference changes in recommender systems via time-dependent Markov random fields. *Expert Syst. Appl.* **2023**, *234*, 121072. [CrossRef]
28. Beheshti, A.; Yakhchi, S.; Mousaeirad, S.; Ghafari, S.M.; Goluguri, S.R.; Edrisi, M.A. Towards Cognitive Recommender Systems. *Algorithms* **2020**, *13*, 176. [CrossRef]
29. Lerner, J.S.; Li, Y.; Valdesolo, P.; Kassam, K.S. Emotion and decision making. *Annu. Rev. Psychol.* **2015**, *66*, 799–823. [CrossRef]
30. Alfaifi, Y.H. Recommender Systems Applications: Data Sources, Features, and Challenges. *Information* **2024**, *15*, 660. [CrossRef]
31. Livne, A.; Unger, M.; Shapira, B.; Rokach, L. Deep Context-Aware Recommender System Utilizing Sequential Latent Context. In *Proceedings of the 13th ACM Conference on Recommender Systems—CARS Workshop, Copenhagen, Denmark; ACM: New York, NY, USA, 2019*. <https://arxiv.org/abs/1909.03999>. [CrossRef]
32. Zhao, W.X.; Guo, Y.; He, Y.; Jiang, H.; Wu, Y.; Li, X. We Know What You Want to Buy: A Demographic-Based System for Product Recommendation on Microblogs. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14), New York, NY, USA; ACM: New York, NY, USA, 2014*; pp. 1935–1944. [CrossRef]
33. Tkalčič, M.; De Carolis, B.; de Gemmis, M.; Odić, A.; Košir, A. *Emotions and Personality in Personalized Services: Models, Evaluation and Applications*; Springer: Berlin/Heidelberg, Germany, 2016. [CrossRef]
34. Ricci, F.; Rokach, L.; Shapira, B. Introduction to recommender systems handbook. In *Recommender Systems Handbook*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 1–35. [CrossRef]
35. Beladev, M.; Rokach, L.; Shapira, B. Recommender Systems for Product Bundling. *Knowl.-Based Syst.* **2016**, *111*, 193–206. [CrossRef]
36. Kouki, P.; Fountalis, I.; Vasiloglou, N.; Yan, N.; Ahsan, U.; Al Jadda, K.; Qu, H. Product Collection Recommendation in Online Retail. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19), Copenhagen, Denmark; ACM: New York, NY, USA, 2019*; pp. 486–490. [CrossRef]
37. Wu, S.; Sun, F.; Zhang, W.; Xie, X.; Cui, B. Graph Neural Networks in Recommender Systems: A Survey. *arXiv* **2020**, arXiv:2011.02260. <https://arxiv.org/abs/2011.02260>. [CrossRef]
38. Zhang, P.; Niu, Z.; Ma, R.; Zhang, F. Multi-view graph contrastive representation learning for bundle recommendation. *Inf. Process. Manag.* **2025**, *62*, 103956. [CrossRef]
39. Dara, S.; Chowdary, C.R.; Kumar, C. A Survey on Group Recommender Systems. *J. Intell. Inf. Syst.* **2019**, *54*, 271–295. [CrossRef]
40. Hwangbo, H.; Kim, Y.S.; Cha, K.J. Recommendation system development for fashion retail e-commerce. *Electron. Commer. Res. Appl.* **2018**, *28*, 94–101. [CrossRef]
41. Kompan, M.; Gaspar, P.; Macina, J.; Cimerman, M.; Bielikova, M. Exploring Customer Price Preference and Product Profit Role in Recommender Systems. *IEEE Intell. Syst.* **2022**, *37*, 89–98. [CrossRef]
42. Shin, D.; Zhong, B.; Biocca, F.A. Beyond user experience: What constitutes algorithmic experiences? *Int. J. Inf. Manag.* **2020**, *52*, 102061. [CrossRef]
43. Wakil, K.; Alyari, F.; Ghasvari, M.; Lesani, Z.; Rajabion, L. A new model for assessing the role of customer behavior history, product classification, and prices on the success of the recommender systems in e-commerce. *Kybernetes* **2020**, *49*, 1325–1346. [CrossRef]
44. Zheng, Y.; Gao, C.; He, X.; Li, Y.; Jin, D. Price-aware recommendation with graph convolutional networks. In *Proceedings of the Proceedings—International Conference on Data Engineering, Dallas, TX, USA, 20–24 April 2020*; pp. 133–144. [CrossRef]
45. Lopatovska, I.; Mokros, H.B. Willingness to pay and experienced utility as measures of affective value of information objects: Users' accounts. *Inf. Process. Manag.* **2008**, *44*, 92–104. [CrossRef]
46. Chen, J.; Jin, Q.; Zhao, S.; Bao, S.; Zhang, L.; Su, Z.; Yu, Y. Does product recommendation meet its Waterloo in unexplored categories? No, price comes to help. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, QLD, Australia, 6–11 July 2014*; pp. 667–676. [CrossRef]

47. Kent, R.J.; Monroe, K.B. The Effects of Framing Price Promotion Messages on Consumers' Perceptions and Purchase Intentions. *J. Retail.* **1998**, *74*, 353–372. [\[CrossRef\]](#)
48. Greenstein-Messica, A.; Rokach, L. Personal price aware multi-seller recommender system: Evidence from eBay. *Knowl.-Based Syst.* **2018**, *150*, 14–26. [\[CrossRef\]](#)
49. Terui, N.; Dahana, W.D. Price customization using price thresholds estimated from scanner panel data. *J. Interact. Mark.* **2006**, *20*, 58–70. [\[CrossRef\]](#)
50. Sato, M.; Izumo, H.; Sonoda, T. Discount sensitive recommender system for retail business. In Proceedings of the EMPIRE '15: 3rd Workshop on Emotions and Personality in Personalized Systems 2015, Vienna, Austria, 16–20 September 2015; pp. 33–40. [\[CrossRef\]](#)
51. Jannach, D.; Adomavicius, G. Price and Profit Awareness in Recommender Systems. *arXiv* **2017**, arXiv:1707.08029. [\[CrossRef\]](#)
52. Sato, M.; Izumo, H.; Sonoda, T. Model of Personal Discount Sensitivity in Recommender Systems. *arXiv* **2016**. Available online: https://ixdea.org/28_6/ (accessed on 28 October 2025).
53. Das, A.; Mathieu, C.; Ricketts, D. Maximizing profit using recommender systems. *arXiv* **2009**, arXiv:0908.3633. [\[CrossRef\]](#)
54. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
55. LiChun, C.; ZhiMin, Z. An overview of deep reinforcement learning. In Proceedings of the 2019 4th International Conference on Automation, Control and Robotics Engineering, Shenzhen, China, 19–21 July 2019. [\[CrossRef\]](#)
56. Wang, C.; Dong, T.; Chen, L.; Zhu, G.; Chen, Y. Multi-objective optimization approach for permanent magnet machine via improved soft actor-critic based on deep reinforcement learning. *Expert Syst. Appl.* **2025**, *264*, 125834. [\[CrossRef\]](#)
57. Xin, X.; Karatzoglou, A.; Arapakis, I.; Jose, J.M. Self-supervised reinforcement learning for recommender systems. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual, 25–30 July 2020; pp. 931–940. [\[CrossRef\]](#)
58. Wu, Y.; MacDonald, C.; Ounis, I. Partially observable reinforcement learning for dialog-based interactive recommendation. In Proceedings of the RecSys 2021—15th ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September–1 October 2021; pp. 241–251. [\[CrossRef\]](#)
59. Luo, W.; Yu, X.; Yen, G.G.; Wei, Y. Deep reinforcement learning-guided coevolutionary algorithm for constrained multiobjective optimization. *Inf. Sci.* **2025**, *692*, 121648. [\[CrossRef\]](#)
60. Lei, Y.; Wang, Z.; Li, W.; Pei, H. Social attentive deep Q-network for recommendation. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 1189–1192. [\[CrossRef\]](#)
61. Farris, L. Optimized recommender systems with deep reinforcement learning. *arXiv* **2021**, arXiv:2110.03039. [\[CrossRef\]](#)
62. Van Hasselt, H.; Guez, A.; Silver, D. Deep reinforcement learning with double Q-learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 2094–2100.
63. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing Atari with deep reinforcement learning. *arXiv* **2013**, arXiv:1312.5602. [\[CrossRef\]](#)
64. Li, J.; Chen, B. A Deep Q-Learning Optimization Framework for Dynamic Pricing in E-Commerce. In *Proceedings of the 2025 4th International Conference on Cyber Security, Artificial Intelligence and the Digital Economy (CSAIDE 2025)*, Kuala Lumpur, Malaysia; ACM: New York, NY, USA, 2025; pp. 367–371. [\[CrossRef\]](#)
65. Fraija, A.; Henao, N.; Agbossou, K.; Kelouwani, S.; Fournier, M.; Nagarsheth, S.H. Deep Reinforcement Learning Based Dynamic Pricing for Demand Response Considering Market and Supply Constraints. *Smart Energy* **2024**, *14*, 100139. [\[CrossRef\]](#)
66. Nomura, Y.; Liu, Z.; Nishi, T. Deep Reinforcement Learning for Dynamic Pricing and Ordering Policies in Perishable Inventory Management. *Appl. Sci.* **2025**, *15*, 2421. [\[CrossRef\]](#)
67. Wang, G.; Ding, J.; Hu, F. Deep Reinforcement Learning Recommendation System Algorithm Based on Multi-Level Attention Mechanisms. *Electronics* **2024**, *13*, 4625. [\[CrossRef\]](#)
68. Rossiiev, O.D.; Shapovalova, N.N.; Rybalchenko, O.H.; Striuk, A.M. A Comprehensive Survey on Reinforcement Learning-based Recommender Systems: State-of-the-Art, Challenges, and Future Perspectives. In Proceedings of the 7th Workshop for Young Scientists in Computer Science and Software Engineering (CSSE@SW 2024). CEUR Workshop Proceedings, Kryvyi Rih, Ukraine, 27 December 2025; pp. 428–440.
69. Gupta, R.; Lin, J.; Meng, F. Multi-agent Deep Reinforcement Learning for Interdependent Pricing in Supply Chains. *arXiv* **2025**, arXiv:2507.02698.
70. Zhang, D.; Zhao, Y.; Sun, L. Reinforcement Learning in Recommender Systems: Progress, Challenges, and Opportunities. In Proceedings of the 18th ACM Conference on Recommender Systems (RecSys), Bari, Italy, 14–18 October 2024; ACM: New York, NY, USA, 2024; pp. 112–129.
71. Dulac-Arnold, G.; Evans, R.; van Hasselt, H.; Sunehag, P.; Lillicrap, T.; Hunt, J.; Mann, T.; Weber, T.; Degris, T.; Coppin, B. Deep Reinforcement Learning in Large Discrete Action Spaces. *arXiv* **2015**, arXiv:1512.07679. [\[CrossRef\]](#)

72. Liu, S.; Cai, Q.; Sun, B.; Wang, Y.; Jiang, J.; Zheng, D.; Gai, K.; Jiang, P.; Zhao, X.; Zhang, Y. Exploration and Regularization of the Latent Action Space in Recommendation. *arXiv* **2023**, arXiv:2302.03431. [[CrossRef](#)]
73. Liu, D.; Li, J.; Wu, J.; Du, B.; Chang, J.; Li, X. Interest Evolution-driven Gated Neighborhood aggregation representation for dynamic recommendation in e-commerce. *Inf. Process. Manag.* **2022**, *59*, 102982. [[CrossRef](#)]
74. Wang, K.; Zou, Z.; Zhao, M.; Deng, Q.; Shang, Y.; Liang, Y.; Wu, R.; Shen, X.; Lyu, T.; Fan, C. RL4rs: A real-world dataset for reinforcement learning based recommender system. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, 23–27 July 2023; pp. 2935–2944.
75. Yu, J.; Xia, X.; Chen, T.; Cui, L.; Hung, N.Q.V.; Yin, H. XSimGCL: Towards Extremely Simple Graph Contrastive Learning for Recommendation. *IEEE Trans. Knowl. Data Eng.* **2023**, *36*, 913–926. [[CrossRef](#)]
76. Wang, W.; Xu, Y.; Feng, F.; Lin, X.; He, X.; Chua, T. Diffusion Recommender Model. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), Taipei, Taiwan, 23–27 July 2023; pp. 832–841. [[CrossRef](#)]
77. Zhao, X.; Xia, L.; Zhang, L.; Ding, Z.; Yin, D.; Tang, J. Deep Reinforcement Learning for Page-wise Recommendations. In Proceedings of the Twelfth ACM Conference on Recommender Systems (RecSys '18), Vancouver, BC, Canada, 2–7 October 2018; pp. 95–103. [[CrossRef](#)]
78. Grandini, M.; Bagli, E.; Visani, G. Metrics for Multi-Class Classification: An Overview. *arXiv* **2020**, arXiv:2008.05756. [[CrossRef](#)]
79. Hand, D.J.; Till, R.J. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Mach. Learn.* **2001**, *45*, 171–186. [[CrossRef](#)]
80. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous Control with Deep Reinforcement Learning. In Proceedings of the International Conference on Learning Representations (ICLR). International Conference on Learning Representations, San Juan, CA, USA, 2–4 May 2016.
81. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 1861–1870.
82. Ng, A.Y.; Russell, S.J. Algorithms for Inverse Reinforcement Learning. In Proceedings of the Seventeenth International Conference on Machine Learning (ICML), Stanford, CA, USA, 29 June–2 July 2000; pp. 663–670.
83. Roijers, D.M.; Vamplew, P.; Whiteson, S.; Dazeley, R. A Survey of Multi-Objective Sequential Decision-Making. *J. Artif. Intell. Res.* **2013**, *48*, 67–113. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.