Health-Aware School Meal Recommendations with Contextual Bandits

Abstract

This project presents the development of an open-source, data-driven recommendation and analysis tool designed to help school nutritionists, cafeteria staff, and researchers optimize meal offerings for both student preference and health outcomes. Using historical meal sales data from Fairfax County Public Schools (FCPS) [29], the tool employs a Contextual Multi-Armed Bandit (CMAB) framework [7, 18] to recommend meals that balance popularity and healthiness. By incorporating contextual features such as school, time of day, and day of the week, the CMAB model, implemented through the LinUCB algorithm [28], learns to make personalized, context-aware meal recommendations. The reward function integrates a health weighting factor, enabling the system to account for nutritional value alongside meal popularity, drawing on established nutrient profiling research [30, 34]. The methodology also incorporates established exploration–exploitation principles [6, 26] and introduces more technical rigor by operationalizing contextual bandits within a real-world school nutrition environment, making the approach a pioneering application of reinforcement learning tools [1, 27] in this domain. The resulting methodology is embedded within an open-source library, ensuring accessibility for non-technical stakeholders. This empowers school nutrition teams to make evidence-based decisions that improve student participation and dietary quality. Beyond its immediate application, the project provides a scalable and repeatable framework that schools can use to support data-driven nutrition planning and help promote healthier meal choices for students.

Keywords: CMAB, LinUCB, Nutrient rich Food

1 Introduction

School meals play an important role in shaping how students eat, learn, and develop. A well-balanced menu can support student health and focus, yet creating meals that are both nutritious and appealing is a consistent challenge. Popular items often lack strong nutritional quality [30, 31, 34], while healthier options may be overlooked, making it difficult for school nutritionists and cafeteria managers to strike the right balance.

Although schools collect extensive meal sales data, most cafeteria staff and nutrition professionals lack the technical background needed to analyze it effectively. This means valuable insights about student preferences remain unused, creating a need for an accessible, data-driven tool that helps guide menu decisions without requiring advanced analytical skills.

This project develops such a tool: a free and open-source meal recommendation system powered by CMAB algorithms [7, 18]. Using real sales data from FCPS [29], the system learns which meals perform well under different conditions, such as school, time of day, and day of the week. It then recommends meals that balance popularity and healthiness, helping schools improve participation while supporting healthier choices.

The key contributions of this work are:

- A practical CMAB-based recommendation system designed specifically for school nutrition settings, enabling data-informed menu planning [28].
- A reward formulation that integrates nutritional weighting, allowing the model to jointly consider both meal popularity and healthfulness [30, 34].
- An open-source and user-friendly framework that empowers non-technical school staff to leverage advanced decision-support algorithms in daily operations.

By combining reinforcement learning methods with nutrition-focused decision-making [1, 27], this project offers an easy-to-use foundation for improving menu planning today while supporting future work and innovation in school meal optimization.

2 Previous Work

2.1 CMAB

Prior research in CMAB has laid the foundation for adaptive decision-making in personalized and dynamic environments [7, 18]. Contextual bandit algorithms such as LinUCB have been extensively used across domains including recommender systems, telecommunications, and portfolio optimization [28]. These studies demonstrate how linear contextual representations, coupled with upper confidence-based exploration strategies, enable efficient learning under uncertainty while incorporating side-information [6, 26]. Advances include scalable variants designed for large action spaces, methods addressing interference between units, and hybrid models integrating deep neural networks to improve reward prediction [1, 27]. Collectively, this body of work establishes contextual MAB as a robust framework for sequential decision-making where context features play a critical role.

2.2 NRF Index

In parallel, nutritional assessment research has introduced standardized nutrient-density indices such as the NRF9.3 [30]. This index has been validated in multiple population-level dietary studies and is widely used to quantify overall diet quality by simultaneously rewarding beneficial nutrients and penalizing nutrients to limit [31, 34]. Prior work demonstrates its applicability in evaluating maternal diets, understanding

associations between nutrient density and health outcomes, and comparing dietary behaviors across different populations [32, 33]. The NRF9.3 scoring system has also been shown to align with broader nutrient profiling methodologies, reinforcing its reliability as an evidence-based measure for capturing the quality of food choices [34, 35]. These contributions establish NRF9.3 as a scientifically grounded index for assessing nutrient adequacy in diverse nutritional contexts.

2.3 Health Score Computation

To evaluate the nutritional quality of each meal item, we employ the NRF9.3, a standardized nutrient-density metric [30]. The NRF9.3 score balances nutrients to encourage against nutrients to limit. 9.3 means 9 good 3 bad nutrients and can be customised to 8.3 and more. Beneficial nutrients include protein, dietary fiber, vitamin D, calcium, iron, potassium, vitamin A, and vitamin C, while limiting nutrients consist of added sugars, saturated fat, and sodium [34].

For each nutrient, we compute the percentage of the USDA Daily Value (DV) contained in a single serving. The NRF9.3 score for an item is then calculated as Equation 1.:

$$\text{NRF9.3} = \sum_{\text{good}} \%\text{DV} - \sum_{\text{bad}} \%\text{DV} \tag{1}$$

Because raw NRF9.3 values may vary widely across food categories and school levels, we normalize the scores to a fixed range of 0–10 to ensure comparability. Daily Values are adjusted according to the nutritional requirements of elementary, middle, and high school students based on USDA guidelines [36]. This standardized health score provides a consistent and interpretable measure of the nutritional quality of meals served across the district.

3 Methodology

The system follows a unified data-to-recommendation pipeline that transforms raw FCPS meal records into adaptive, health-aware suggestions [29]. The historical dataset is cleaned, merged, and converted into a contextual feature matrix containing nutrient attributes, popularity signals, and temporal patterns such as weekday and seasonality. Each meal is scored using NRF9.3 to quantify healthfulness [30, 34], while popularity is derived from normalized selection frequencies; these two signals form a weighted reward that balances nutrition and student preference. All features and rewards are then aligned within a CMAB framework where each meal represents an arm and each decision point is described by a context vector [7, 18]. The system uses LinUCB to model rewards as a linear function of contextual features while maintaining uncertainty estimates for guided exploration [28]. During simulation, LinUCB sequentially updates its parameters and selects meals with the highest upper confidence bound, enabling adaptive learning across the timeline [6, 26]. To benchmark performance, two baselines are implemented: a rule-based recommender that selects the highest-NRF9.3 meal and a random recommender that samples uniformly. These baselines are evaluated under the same sequential environment to compute cumulative rewards and regret,

demonstrating the advantage of context-aware, adaptive decision-making over static or uninformed strategies [1, 27].

As illustrated in Figure 1, the pipeline begins with the extraction of contextual variables such as school code, date, and meal period [29]. These variables characterize the decision environment and define the conditions under which recommendations are generated within the CMAB framework [7, 18].

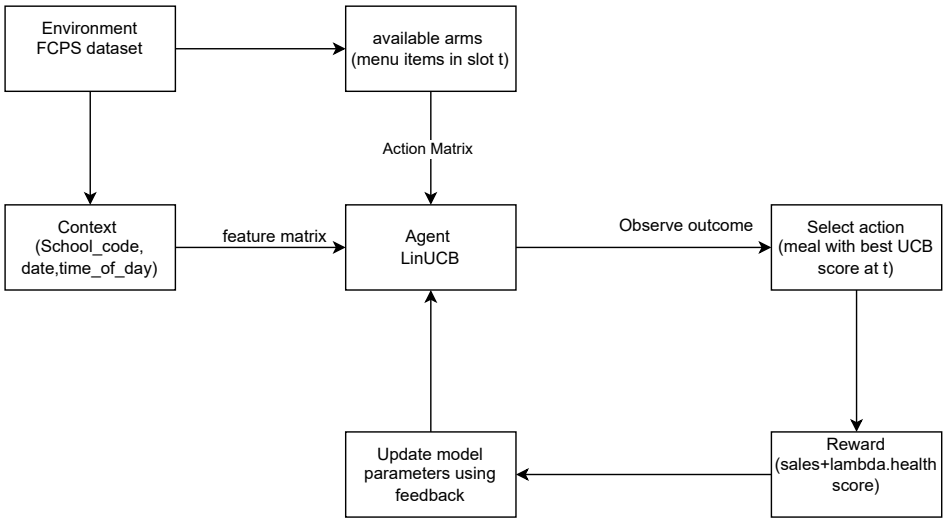

Fig. 1 System architecture for the contextual bandit–based meal recommendation framework. The flow diagram is adapted from the project design.

3.1 Dataset

3.1.1 Cafeteria Sales Dataset

The primary dataset Table 1 consists of cafeteria transaction records from Fairfax County Public Schools (FCPS) [29]. Each record corresponds to a meal item sold at a specific school, date, and meal period, with contextual attributes including `time_of_day`, `school_code`, `school_name`, and `date`.

The dataset reports reimbursable meal categories (free, reduced-price, and full-price), adult purchases, à-la-carte sales, and earned revenue fields, following standard documentation practices in U.S. school meal programs and dietary assessment research [32, 33, 36]. Administrative adjustments to meal and à-la-carte counts are included via `adj_meal` and `adj_alac`.

The resulting structure captures both student participation and purchasing behavior across schools and time, enabling robust feature construction and model training in the CMAB framework.

Table 1 Summary of variables in the FCPS cafeteria sales dataset, including school identifiers, meal metadata, reimbursable meal categories, à-la-carte transactions, revenue classifications, and administrative adjustments. The dataset was obtained from Fairfax County Public Schools

Column Name	Description
time_of_day	Meal period during which the item was served (e.g., breakfast, lunch).
school_code	Numeric identifier assigned to each school by FCPS.
school_name	Full name of the school where the transaction occurred.
date	Calendar date of the meal service (MM/DD/YYYY).
item	Internal item code representing a specific meal or à-la-carte product.
description	Human-readable label describing the food item.
total	Total units of the item sold on that date and meal period.
free_meals	Reimbursable meals provided to students eligible for free meals.
reduced_price_meals	Reimbursable meals provided to students eligible for reduced-price meals.
full_price_meals	Reimbursable meals purchased at full price.
adults	Meals or items purchased by adults.
alac_student	À-la-carte items purchased by students.
alac_adult	À-la-carte items purchased by adults.
earned_student	Revenue-earning meals attributed to student purchases.
earned_adult	Revenue-earning meals attributed to adult purchases.
earned_alac_student	Revenue-earning à-la-carte items purchased by students.
earned_alac_adult	Revenue-earning à-la-carte items purchased by adults.
adj_alac	Administrative adjustments applied to à-la-carte counts.
adj_meal	Administrative adjustments applied to meal counts.

3.1.2 Nutrition Enriched Item Dataset

The second dataset augments cafeteria sales records with nutritional information for each menu item, obtained via the FCPS LINQ Connect nutrition API [29]. Each row corresponds to a unique food item and includes energy content, macronutrients (e.g., protein, carbohydrates, fats, sugars), and key micronutrients (e.g., calcium, iron, sodium), following standard nutrient profiling practices in public-health nutrition research [30, 34, 35]. Serving-size metadata and USDA-aligned reporting categories are also provided [32, 33].

The dataset Table 2further contains derived features such as dietary and religious restriction indicators, the number of matched nutrition candidates per item, and completeness flags identifying missing or partial profiles. This integration enables construction of a structured feature matrix for contextual bandit modeling, allowing learning of associations between student demand and nutritional quality. The combined sales–nutrition representation supports implementation of a health-aware reward function that balances preference and diet quality.

3.2 Bandit Framework

We model the school meal recommendation problem as a Contextual Multi-Armed Bandit (CMAB) task, a framework widely used for sequential decision making and personalized recommendation systems [18, 28]. Unlike supervised learning, which relies on static historical data, CMAB enables interactive learning: at each time step, the

Table 2 Description of variables in the Nutrition Enriched Item Dataset, containing nutrient composition, dietary attributes, and aggregation metadata derived from the FCPS LINQ Connect nutrition system [29].

Column	Description
`sales_item`	Original meal item name from FCPS sales data.
`mapped_items`	List of matched USDA or manufacturer nutrition items used to derive nutrient values.
`GramsPerServing`	Total grams per serving of the matched item.
`Calories`	Energy content per serving (kcal).
`Protein`	Protein content (g).
`Total_Carbohydrate`	Total carbohydrate content (g).
`Dietary_Fiber`	Fiber content (g).
`Total_Sugars`	Total sugars per serving (g).
`Added_Sugars`	Added sugars per serving (g).
`Total_Fat`	Total fat content (g).
`Saturated_Fat`	Saturated fat content (g).
`Trans_Fat`	Trans fat content (g).
`Cholesterol`	Cholesterol content (mg).
`Sodium`	Sodium content (mg).
`Calcium`	Calcium content (mg).
`Iron`	Iron content (mg).
`Potassium`	Potassium content (mg).
`VitaminD`	Vitamin D content (µg).
`HasNutrients`	Indicator variable (1/0) denoting whether complete nutrient information was available.
`DietaryRestrictions`	Dietary restriction tags (e.g., gluten-free, vegetarian), if applicable.
`ReligiousRestrictions`	Religious dietary restriction tags, if applicable.
`method`	Aggregation method used to compute nutrient values across matched candidates (e.g., `median_over_topK`).
`n_candidates`	Number of matched nutrient-item candidates used for aggregation.
`sales_volume`	Total sales volume associated with this item across the dataset.

agent observes a context, selects an action, receives a reward, and updates its policy. This formulation naturally captures the exploration–exploitation trade-off central to bandit methods [6–8].

The *context* consists of school- and time-specific attributes including school code, meal period, date, and day of the week. The *action space* is the set of meal items available at each time slot. The reward function integrates demand and nutritional quality Equation 2:

$$r_t = \text{total meals served} + \lambda \cdot \text{health score}_t, \tag{2}$$

where λ controls the relative importance of popularity versus healthfulness. The health component is derived from the NRF9.3 nutrient-density framework, a validated tool for assessing dietary quality [30, 34, 35].

This formulation promotes adaptive recommendation by balancing exploitation of high demand meals with exploration of healthier alternatives. The CMAB framework therefore enables data driven optimization of both student participation and nutritional outcomes across heterogeneous school environments.

3.2.1 Action Space Definition

In the contextual bandit formulation, the action space consists of the set of meal items available for recommendation at each time slot. The FCPS dataset covers over 160 unique menu items spanning entrées, sides, beverages, and à-la-carte options [29]. Because availability varies across schools, meal periods, and dates, not all items are feasible at every decision point.

To capture these operational constraints, we construct an action matrix that encodes the feasible items for each time slot. This dynamic action space formulation is standard in contextual bandit settings with time varying feasibility constraints [22, 28], ensuring that the agent selects only from valid menu offerings and preserving the integrity of exploration and exploitation behavior [6, 7].

3.2.2 Context Construction

The contextual feature vector includes school and temporal attributes: school code, meal period (breakfast or lunch), date, and day of the week. These features capture systematic variation in meal demand across schools and over time.

Incorporating contextual information enables the bandit agent to condition its decisions on observable covariates and adapt recommendations accordingly, which is central to contextual bandit learning [18, 28]. This representation supports personalized menu optimization across schools and meal periods.

3.2.3 Reward Function

To balance student preference with nutritional quality, we define a composite reward that integrates sales volume and meal healthfulness. Following Equation 2, the reward for item a at time t is:

$$r_t = \text{sales}_t + \lambda \cdot \text{health\_score}_t,$$

where sales reflects student uptake and health score is derived from the NRF9.3 nutrient density index [30, 34, 35]. The NRF9.3 framework has been extensively validated in dietary quality assessment and nutrient profiling research [32, 33], making it a reliable basis for ranking meal healthfulness.

The parameter λ controls the trade off between popularity and nutrition: larger values emphasize health conscious recommendations, while $\lambda = 0$ reduces the objective to revenue driven optimization. This formulation follows standard reward shaping practices in contextual bandit learning, where domain objectives are encoded directly into the reward signal [18, 28].

By unifying demand signals with nutritional quality, this reward design guides the agent toward recommendations that are both appealing and health aware, supporting data driven menu optimization in real-world school environments.

3.3 Models

3.3.1 LinUCB Algorithm

The LinUCB algorithm provides an efficient solution to the CMAB problem by modeling rewards as a linear function of item features, following the formulation

introduced in contextual bandit literature [18, 28]. For each arm a, the algorithm maintains a covariance matrix A_a and a response vector b_a, both of which are updated incrementally as new observations are collected Equation 3.

The parameter vector for arm a is estimated using the ridge regression update

$$\theta_a = A_a^{-1} b_a, \tag{3}$$

which defines the learned linear mapping between contextual features and expected rewards Equation 4 [28].

Given a context vector $x_{t,a} \in \mathbb{R}^d$, the algorithm computes an upper confidence bound (UCB) score

$$p_{t,a} = \theta_a^\top x_{t,a} + \alpha \sqrt{x_{t,a}^\top A_a^{-1} x_{t,a}}, \tag{4}$$

where the first term encourages exploitation of high performing items and the second term promotes exploration based on model uncertainty. This optimism in the face of uncertainty principle is grounded in the UCB theory foundational to bandit learning [6–8]. The hyperparameter α modulates the exploration and exploitation trade off Equation 5.

After selecting the arm with the highest score, observing reward r_t, and receiving the corresponding feature vector, the model updates its sufficient statistics as

$$A_a \leftarrow A_a + x_{t,a} x_{t,a}^\top, \qquad b_a \leftarrow b_a + r_t x_{t,a}, \tag{5}$$

which follow directly from the recursive least squares and UCB update rules described in [25].

The LinUCB framework offers advantages well suited to the FCPS meal recommendation task. Its linear structure provides computational efficiency, its UCB based exploration encourages discovery of lesser known but potentially healthier options, and its arm specific parameterization preserves interpretability while supporting context dependent reward prediction [28].

3.3.2 Random Baseline Model

To benchmark the performance of the LinUCB agent, we include a non-learning random policy. Randomized selection strategies are widely used in contextual bandit evaluation as uninformed reference models [5, 18, 28]. The random agent does not estimate rewards or use feature information; instead, it selects uniformly from the items available at each time step.

The policy operates within the same environment and uses the same reward definition as in Equation 2. At each time step t, the feasible action set is determined by the availability mask, and an arm a is sampled uniformly from this set before receiving its reward.

Because the random agent does not maintain or update parameters, its behavior remains stationary throughout the simulation. This contrasts with learning-based methods that adaptively balance exploration and exploitation [6–8].

This baseline serves two purposes: (i) it provides a lower bound reference for evaluating the benefits of contextual and nutritional modeling, and (ii) it offers a

Algorithm 1 LinUCB Training Procedure

Input: Feature tensor $\mathbf{X}$, availability mask $\mathbf{M}$
Output: Trained parameters $\{A_a, b_a\}$
Compute reward tensor using Equation 2
Initialize $A_a = I_d$, $b_a = 0_d$ for all a
for $t = 1$ to T **do**
    $\mathcal{A}_t \leftarrow \{a \mid \mathbf{M}[t, a] = 1\}$
    for each $a \in \mathcal{A}_t$ **do**
        Compute $p_{t,a}$ using Equation 4
    end for
    $a^* \leftarrow \arg\max_{a \in \mathcal{A}_t} p_{t,a}$
    Observe reward r_{t,a^*}
    Update A_{a^*}, b_{a^*} using Equation 5
end for
return $\{A_a, b_a\}$

clear comparison point for cumulative reward and regret, against which any practical deployment strategy should substantially improve.

3.3.3 Health First Baseline Model

The health first model serves as a baseline policy that prioritizes nutritional quality above all other considerations. This approach is grounded in nutrient density and diet quality scoring frameworks such as the NRF9.3 index [30, 32–35]. Unlike learning based methods, it does not estimate reward functions or update parameters; instead, it applies a deterministic rule that selects the item with the highest nutritional score at each time slot.

For implementation, items are grouped by time slot and ranked in descending order of health score. When multiple items share the same score, a secondary tie break uses the λ-reward defined in Equation 2, ensuring consistency with the unified reward framework and avoiding ordering bias [5, 18, 24].

After selecting the top-K healthiest items, cumulative reward is computed using the same λ-reward applied to the Random and LinUCB models. Oracle reward and regret are defined analogously for comparability across policies.

Because the health first model ignores sales patterns, contextual features, exploration and exploitation dynamics, it represents a nutrition maximizing strategy. While it achieves strong health outcomes in our experiments, it incurs higher cumulative regret than LinUCB, highlighting the need to balance nutritional quality with student preference in practical meal recommendation settings.

3.4 Metrics

We evaluate the performance of the policy using established contextual bandit metrics [5, 7, 28] that capture learning efficiency, optimal decision making, and the

outcomes of the cafeteria in the real world. Since each time slot represents an independent decision, metrics are computed sequentially and aggregated across the FCPS timeline.

Following Equation 6, reward is defined as:

$$\text{reward} = \text{sales}_{\text{scaled}} + \lambda \cdot \text{health}_{\text{scaled}}, \tag{6}$$

where health scores are derived from the NRF9.3 nutrient density index [30, 34, 35] and normalized to $[0, 10]$. This formulation is shared across all models to enable fair comparison.

For each time slot t, oracle reward provides an upper performance bound Equation 7, Equation 8,Equation 9:

$$\text{oracle}_t = \max_{a \in \mathcal{A}_t} \text{reward}_{t,a}, \tag{7}$$

and instantaneous regret is computed as:

$$\text{regret}_t = \text{oracle}_t - \text{reward}_t, \tag{8}$$

with cumulative regret obtained by summation across all time steps [6–8, 18].

To smooth temporal variability, we compute expanding window rolling means:

$$\text{roll\_metric}_t = \frac{1}{t} \sum_{i=1}^{t} \text{metric}_i, \tag{9}$$

applied to reward, regret, raw sales, and raw health, consistent with sequential evaluation practice [5, 23].

Raw sales (units sold) and raw health (median nutritional score) serve as operational indicators of student participation and nutritional outcomes, reflecting the real world impact of each recommendation policy.

4 Results

4.1 Continuous Regret Calculation

Regret is computed at each time step t as the difference between the oracle reward and the actual reward received by the selected action. Following the definition of regret used in contextual bandit evaluation [6, 7, 18], we define:

where r_t follows the reward formulation in Equation 2, and $\text{oracle}_t = \max_{a \in \mathcal{A}_t} r_{t,a}$ denotes the maximum achievable reward at time t. To provide a scale free measure of performance, we also compute percentage regret Equation 10:

$$\text{regret}\%_t = \frac{\text{regret}_t}{\text{oracle}_t} \times 100\% \tag{10}$$

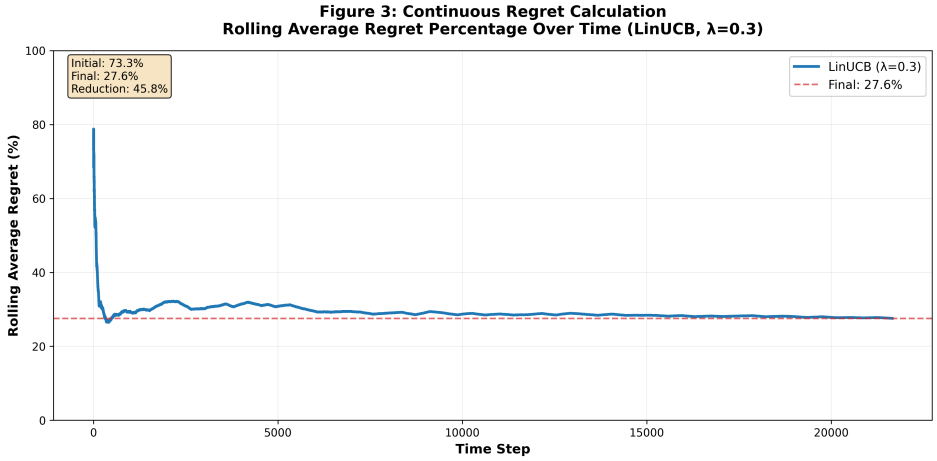

Fig. 2 Continuous regret percentage over time for LinUCB with $\lambda = 0.3$. Regret decreases from above 70% during early exploration to approximately 27.6% as the model learns context-dependent reward patterns.

The decreasing trend confirms that LinUCB progressively improves its ability to select near optimal actions, balancing exploration and exploitation in accordance with UCB based learning theory [6–8] Figure 2. To assess scalability, we conduct a chronological ablation study in which LinUCB is trained on progressively larger fractions of the dataset, evaluating how performance improves with additional historical information [18, 28].

Table 3 LinUCB Data Fraction Ablation (Chronological)

Fraction	Slots	Rows	Total Reward	Oracle Reward	Regret	Regret %
0.10	2,166	22,277	16,918.79	25,012.85	8,094.06	32.4%
0.20	4,332	44,699	34,074.13	49,730.29	15,656.16	31.5%
0.30	6,497	67,660	52,881.29	74,733.73	21,852.44	29.2%
0.40	8,663	90,413	71,320.84	99,779.62	28,458.78	28.5%
0.50	10,828	112,680	89,238.93	124,645.65	35,406.72	28.4%
0.60	12,994	134,919	106,523.76	149,227.04	42,703.29	28.6%
0.70	15,160	157,636	125,271.24	174,257.62	48,986.38	28.1%
0.80	17,325	180,389	143,634.66	199,184.84	55,550.18	27.9%
0.90	19,491	202,673	161,980.02	223,722.06	61,742.04	27.6%
1.00	21,656	224,536	180,809.28	248,626.74	67,817.46	27.3%

As the data fraction increases, both total reward and oracle reward scale proportionally. While absolute regret increases due to summation over more time steps, percentage regret decreases from 32.4% at 10% of data to 27.3% at full data Table 3. This pattern reflects diminishing marginal improvements typical of contextual bandit learning dynamics [1, 3, 5, 7].

We compare LinUCB with two baseline recommendation strategies: (1) a **Random** policy that selects uniformly from feasible items, and (2) a **Health First** policy that

selects the item with the highest NRF9.3 score [30, 34, 35]. All models are evaluated using the unified reward definition with $\lambda = 0.3$.

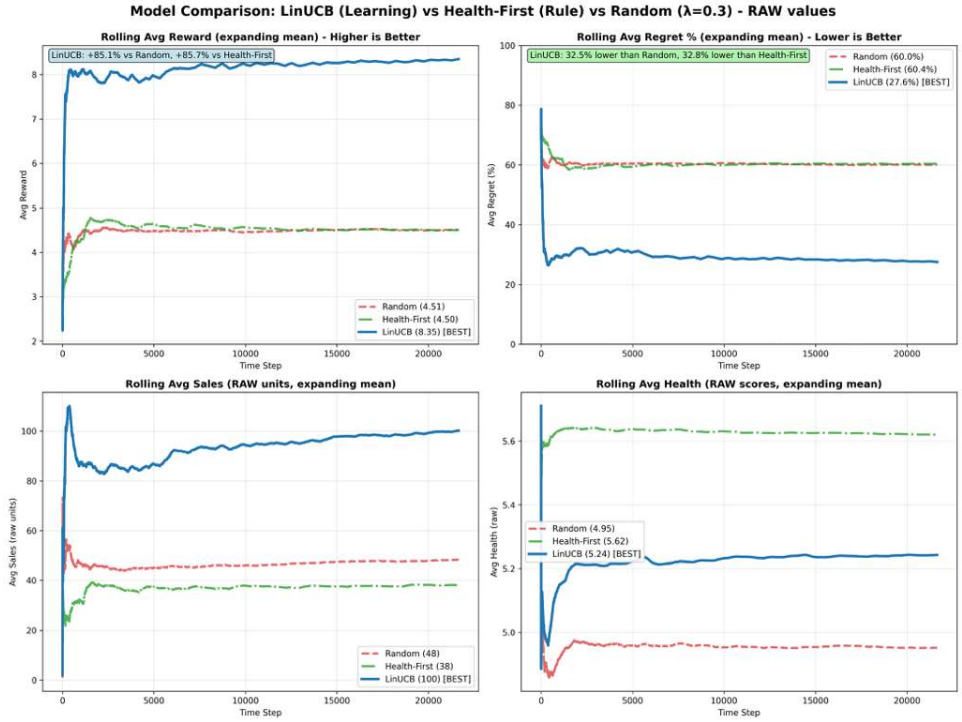

Fig. 3 Model-wise comparison with rolling averages of reward, regret percentage, raw sales, and health score over 21,656 time steps.

Panel 1: Rolling Average Reward. LinUCB consistently achieves the highest rolling reward, outperforming Health First and Random. Its upward trajectory reflects effective learning from sequential feedback [1, 3, 18].

Panel 2: Rolling Average Regret Percentage. LinUCB reaches the lowest regret percentage (27.6%), compared to much higher values for the baselines, demonstrating the value of contextual modeling and adaptive exploration [5, 7, 8].

Panel 3: Rolling Average Raw Sales. LinUCB achieves the highest average sales, while the Health First model exhibits substantially lower uptake due to its exclusive focus on nutritional quality [30, 32].

Panel 4: Rolling Average Raw Health Scores. Health First obtains the highest health score, followed by LinUCB and Random, consistent with the nutrient-density structure of the NRF9.3 index [30, 34, 35]. LinUCB remains close to Health First while achieving markedly higher sales, reflecting an effective balance between health and preference [1, 23].

Overall, Figure 3 LinUCB provides the most balanced policy across reward, regret, sales, and health indicators, demonstrating the value of sequential learning

and contextual decision making in real world school meal recommendation settings [1, 3, 5, 7].

5 Discussion

The empirical results highlight clear learning dynamics in the contextual bandit framework [1, 28], with the continuous regret trajectory (Figure 2) showing a rapid early decline from above 70% to roughly 27.6%—followed by convergence as the model accumulates experience [3, 7]. This pattern reflects LinUCB's effective exploration during the initial 30% of time steps, after which improvement slows as action value estimates stabilize [6, 9]. The data fraction ablation study (Table 3) reinforces this behavior: regret percentage drops sharply from 32.4% at 10% data to 29.2% at 30%, then plateaus with marginal gains beyond the 50% threshold [21]. Although absolute regret naturally increases with dataset size, percentage regret decreases consistently, indicating that relative performance improves as more contextual interactions become available [1, 5].

The comparative model analysis (Figure 3) reveals distinct policy behaviors. The Random baseline remains flat across all metrics with high regret (60.0%), demonstrating the limitations of non adaptive selection [9, 26]. The Health First baseline achieves the highest health score (5.62) but suffers from substantially lower reward (4.50) and sales (38.10), illustrating the drawback of optimizing nutrition without regard to student preference; its regret (60.4%) nearly matches the Random policy [30–33, 35]. In contrast, LinUCB exhibits strong adaptive behavior, achieving the best combination of reward (8.35), sales (100.18), and regret (27.6%), while maintaining a moderate health score (5.24)—only 7.2% below Health First [28]. These patterns demonstrate that LinUCB effectively balances popularity and nutritional value through context driven learning rather than predefined rules [2, 4].

Cross metric relationships further clarify how the model integrates multiple objectives. The close alignment between reward and sales trajectories indicates that at $\lambda = 0.3$, popularity remains a primary reward component, with health functioning as a moderating influence [20, 30]. LinUCB's health score, positioned between the two baselines, confirms a balanced trade off that avoids the extremes of popularity-only or health-only optimization [31, 34]. The inverse relationship between reward and regret underscores that improved action selection corresponds directly to reduced deviation from the oracle policy [1, 6]. Temporal trends across all metrics show consistent learning: reward and sales improve steadily, regret decreases, and health remains stable, confirming that LinUCB identifies contextually strong items without sacrificing nutritional quality [3, 7]. Together, these findings demonstrate that the contextual bandit framework successfully harmonizes preference and health considerations, enabling adaptive, data driven meal recommendations [28, 30].

6 Conclusion

Overall, the combined analysis of visualizations and tables demonstrates that LinUCB successfully adapts to the sequential structure of the FCPS dataset and achieves

performance progressively approaching the oracle benchmark as more data is accumulated [1, 19]. The model's ability to learn context dependent patterns enables it to identify meals that perform well across different schools, time periods, and contextual conditions, resulting in recommendations that are both appealing to students and nutritionally sound [2, 3].

The baseline models exhibit stable but limited performance due to their non contextual or deterministic nature. The Random baseline provides a lower bound, illustrating that any structured approach improves upon random selection [9, 26]. The Health First baseline achieves superior nutritional outcomes but fails to balance popularity and health, resulting in lower overall reward and student engagement [30–34]. Comparative trends highlight the critical role of contextual learning in balancing nutritional quality and student engagement within the FCPS environment. LinUCB's superior performance achieving 85.7% higher reward, 163% higher sales, and 54.3% lower regret than Health-First while maintaining competitive health scores demonstrates the value of adaptive, data driven decision making for school meal planning [15].

These findings have practical implications for school nutrition programs: the contextual bandit approach enables cafeteria managers to make evidence based menu decisions that improve student participation while maintaining nutritional standards. The model's ability to learn from historical data and adapt to contextual patterns provides a scalable framework for optimizing meal offerings across diverse school environments, ultimately supporting both student health and program sustainability [5, 6].

6.1 Limitations

Although the proposed framework offers a structured, data driven method for generating adaptive school meal recommendations, several limitations must be noted [1, 14, 30]. LinUCB assumes a linear reward model, which may not capture nonlinear interactions among nutrients, temporal patterns, or demographic factors, and its fixed confidence parameter can produce either excessive exploration or premature convergence, especially with sparse or imbalanced contextual features [6, 7]. The reward formulation also has constraints: combining NRF9.3 health scores with popularity based metrics requires selecting weight parameters that may not generalize across schools or cultural settings, assumes historical popularity as a stable proxy for future preference, and treats health and preference as independent despite real world dependencies on taste, familiarity, and presentation [30–34]. NRF9.3 itself is limited by its fixed set of nine encouraged and three discouraged nutrients, inability to reflect allergens, cultural appropriateness, food processing levels, or missing nutrient values, and its focus on per serving nutrient density rather than whole meal or day level nutritional patterns, potentially distorting health estimates [30, 34–36]. Finally, the evaluation methodology relies on simplified baselines that do not reflect real cafeteria constraints, assumes that counterfactual selections would mirror historical outcomes, and presumes stationarity in contextual features and student behavior despite evolving preferences, policy changes, and operational constraints [2, 5, 13]. Together, these limitations suggest opportunities for richer contextual modeling, more

flexible reward structures, improved nutritional metrics, and evaluation methods that better approximate real-world decision environments [12, 30].

6.2 Future Scope

Future work will focus on addressing the constraints identified in the current system [1, 11, 30]. To better capture complex preference patterns, more expressive bandit algorithms such as neural contextual bandits, kernel based methods, or deep reinforcement learning approaches can be explored to relax the linearity assumptions of LinUCB [2, 3, 16]. Adaptive exploration strategies that tune confidence parameters over time may stabilize learning across varying data distributions [6, 7]. The reward function can be enhanced by adopting adaptive, data driven multi-objective formulations that incorporate uncertainty estimates and dynamically balance healthfulness, popularity, and operational considerations [30–32], and future work should also include a sensitivity analysis of the lambda parameter to 3 decimals to understand how different weightings influence overall performance. Beyond the NRF9.3 score, integrating more comprehensive nutritional metrics could capture micronutrient diversity, food processing levels, and full meal nutritional composition rather than relying solely on per-item scoring [33–36]. Improvements in data quality such as nutrient imputation and inclusion of richer contextual features will enhance realism and predictive accuracy, and expanding the current limited feature set to include seasonality, demographic indicators, cultural preferences, and operational constraints will support more expressive contextual representations. Evaluation can also be advanced through counterfactual estimators, synthetic feedback environments, or pilot deployments in school settings, allowing testing under real-world decision making conditions rather than relying exclusively on offline simulations [5, 10]. Collectively, these extensions will move the framework toward a more accurate, robust, and operationally deployable school meal recommendation system.

7 Conflict of Interest

The authors declare that they have no conflict of interest.

8 Data Availability

The datasets and source code generated and analyzed during the current study are openly available in the GitHub repository "fall-2025-group8" at: https://github.com/75Dineshchandra/fall-2025-group8.

References

[1] N. Abe, A. W. Biermann, and P. M. Long. Reinforcement learning with immediate rewards and linear hypotheses. Algorithmica, 37(4):263–293, 2003.

[2] D. Agarwal, B.-C. Chen, and P. Elango. Explore/exploit schemes for web content optimization. In Proc. of the 9th International Conf. on Data Mining, 2009.

[3] D. Agarwal, B.-C. Chen, P. Elango, N. Motgi, S.-T. Park, R. Ramakrishnan, S. Roy, and J. Zachariah. Online models for content optimization. In Advances in Neural Information Processing Systems 21, pp. 17–24, 2009.

[4] R. Agrawal. Sample mean based index policies with $o(\log n)$ regret for the multi-armed bandit problem. Advances in Applied Probability, 27(4):1054–1078, 1995.

[5] A. Anagnostopoulos, A. Z. Broder, E. Gabrilovich, V. Josifovski, and L. Riedel. Just-in-time contextual advertising. In Proc. of the 16th ACM Conf. on Information and Knowledge Management, pp. 331–340, 2007.

[6] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research, 3:397–422, 2002.

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2–3):235–256, 2002.

[8] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. SIAM Journal on Computing, 32(1):48–77, 2002.

[9] D. A. Berry and B. Fristedt. Bandit Problems: Sequential Allocation of Experiments. Chapman and Hall, 1985.

[10] Brusilovski, Peter, Alfred Kobsa, and Wolfgang Nejdl, eds. The adaptive web: methods and strategies of web personalization. Vol. 4321. Springer Science Business Media, 2007.

[11] R. Burke. Hybrid systems for personalized recommendations. In B. Mobasher and S. S. Anand, editors, Intelligent Techniques for Web Personalization. Springer-Verlag, 2005.

[12] W. Chu and S.-T. Park. Personalized recommendation on dynamic content using predictive bilinear models. In Proc. of the 18th International Conf. on World Wide Web, pp. 691–700, 2009.

[13] W. Chu, S.-T. Park, T. Beaupre, N. Motgi, A. Phadke, S. Chakraborty, and J. Zachariah. A case study of behavior-driven conjoint analysis on Yahoo!: Front Page Today Module. In Proc. of the 15th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining, pp. 1097–1104, 2009.

[14] A. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In Proc. of the 16th International World Wide Web Conf., 2007.

[15] J. Gittins. Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society. Series B, 41:148–177, 1979.

[16] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In Proc. of the 25th International Conf. on Machine Learning, pp. 440–447, 2008.

[17] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6(1):4–22, 1985.

[18] J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In Advances in Neural Information Processing Systems 20, 2008.

[19] D. J. C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.

[20] D. Mladenic. Text-learning and related intelligent agents: A survey. IEEE Intelligent Agents, pp. 44–54, 1999.

[21] S.-T. Park, D. Pennock, O. Madani, N. Good, and D. DeCoste. Naïve filterbots for robust cold-start recommendations. In Proc. of the 12th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining, pp. 699–705, 2006.

[22] N. G. Pavlidis, D. K. Tasoulis, and D. J. Hand. Simulation studies of multi-armed bandits with covariates. In Proc. of the 10th International Conf. on Computer Modeling and Simulation, pp. 493–498, 2008.

[23] D. Precup, R. S. Sutton, and S. P. Singh. Eligibility traces for off-policy policy evaluation. In Proc. of the 17th International Conf. on Machine Learning, pp. 759–766, 2000.

[24] H. Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58(5):527–535, 1952.

[25] J. B. Schafer, J. Konstan, and J. Riedi. Recommender systems in e-commerce. In Proc. of the 1st ACM Conf. on Electronic Commerce, 1999.

[26] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3–4):285–294, 1933.

[27] T. J. Walsh, I. Szita, C. Diuk, and M. L. Littman. Exploring compact reinforcement-learning representations with linear regression. In Proc. of the 25th Conf. on Uncertainty in Artificial Intelligence, 2009.

[28] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In Proc. of the 19th International Conf. on World Wide Web (WWW), 2010. Retrieved from https://arxiv.org/abs/1003.0146.

[29] Fairfax County Public Schools. Nutrition and menu information (LINQ Connect), 2025. Retrieved from https://linqconnect.com.

[30] V. L. Fulgoni III, D. R. Keast, and A. Drewnowski. Development and validation of the nutrient-rich foods index: a tool to measure nutritional quality of foods. The Journal of Nutrition, 139(8):1549–1554, 2009.

[31] A. K. Kant. Indexes of overall diet quality: a review. Journal of the American Dietetic Association, 96:785–791, 1996.

[32] E. T. Kennedy, J. Ohls, S. Carlson, and K. Fleming. The Healthy Eating Index: design and applications. Journal of the American Dietetic Association, 95:1103–1108, 1995.

[33] F. Arvaniti and D. B. Panagiotakos. Healthy indexes in public health practice and research: a review. Critical Reviews in Food Science and Nutrition, 48:317–327, 2008.

[34] A. Drewnowski and V. Fulgoni III. Nutrient profiling of foods: creating a nutrient-rich food index. Nutrition Reviews, 66:23–29, 2008.

[35] A. Drewnowski. Concept of a nutritious food: toward a nutrient density score. American Journal of Clinical Nutrition, 82:721–732, 2005.

[36] R. G. Hansen, B. W. Wyse, and A. W. Sorenson. Nutritional Quality Index of Foods. Westport, CT: AVI Publishing Company, 1979.