Reducing Institutional Food Waste via Contextual Bandits

Neeraj Shashikant Magadum Author[1,2*]
and Varun Gholap Author[2,3†]

[1*]Department, Organization, Street, City, 100190, State, Country.

*Corresponding author(s). E-mail(s):
neerajshashikant.magadum@gwmail.gwu.edu;
Contributing authors: varun.gholap@gwmail.gwu.edu ;
†These authors contributed equally to this work.

Abstract

This paper introduces a recommendation system for school meal planning using a Contextual Multi-Armed Bandit (CMAB) algorithm. The primary goal is to reduce food waste in school meal programs by improving demand forecasting and aligning production with actual student preferences. Our system uses historical sales data from Fairfax County Public Schools (FCPS) to learn how meal choices vary across different contexts, such as the specific school and day of the week. It recommends meals that are more likely to be selected, helping kitchens plan production quantities more accurately and avoid overproducing unpopular items. We implement a LinUCB algorithm that adapts to changing preferences by balancing familiar options with occasional exploration of less-tested items. We develop the system as part of an open-source tool to support data-driven decision-making and promote more efficient, cost-effective, and sustainable meal operations

Keywords: School meal programs, Food waste reduction, Contextual bandit, Demand forecasting

1 Introduction

School nutrition programs generate large amounts of avoidable food waste due to the inherent difficulty of predicting daily student meal preferences.[4–6] Even small mismatches between planned production and actual participation can lead to substantial overproduction, unnecessary costs, and increased environmental impact. Many districts lack analytical tools that can translate historical meal data into actionable insights, leaving menu planning reliant on intuition and static assumptions rather than evidence-based forecasting.[6]

To address this multi-faceted problem, we propose a data-driven recommendation system based on a Contextual Multi-Armed Bandit (CMAB). This reinforcement learning approach is ideal for navigating the complex trade-offs in menu planning. By learning from the daily feedback of student choices, the system can identify which meals are likely to be successful in a specific context (exploitation) while still exploring less-tested items when useful (exploration).[2, 3, 8] Crucially, the predictive power of the model also serves as a powerful tool for demand forecasting. By anticipating which meals will be chosen, schools can align production quantities with actual demand, directly tackling the costly and wasteful issue of overproduction.

Our model is developed using historical meal sales data from Fairfax County Public Schools (FCPS). In this framework, each food item is an "arm" that can be recommended based on contextual factors like the school, time of day, and date. The system's reward function is designed to primarily optimize for student consumption, since accurately estimating selection likelihood helps reduce waste by preventing overproduction of unpopular items.

The main contribution of this project is the design and application of a CMAB system to address waste reduction in a real-world school meal setting. We implement a LinUCB algorithm and demonstrate its potential for improving demand estimation, guiding menu recommendations, and reducing unnecessary food waste.[2, 3] The system is intended for integration into a free, open-source tool to support school food service teams in making data-driven, efficient, and sustainable operational decisions.

2 Previous Work

Our research constitutes a novel synthesis of methods from recommendation systems, reinforcement learning, and public health informatics, applied to the operational challenges of institutional food service.

Conventional food recommendation systems, which typically use collaborative or content-based filtering to mirror user taste,[1] are fundamentally misaligned with our objectives. Their paradigm of simple preference-matching is inadequate for a setting that must balance student satisfaction against the competing institutional goals of minimizing waste. Our problem is not one of unconstrained preference satisfaction, but of multi-objective optimization.

This dynamic challenge, which requires balancing known favorites (exploitation) with the strategic testing of new items (exploration), necessitates a reinforcement learning approach. We therefore employ a Contextual Multi-Armed Bandit (CMAB) framework, a methodology pioneered in domains like computational advertising for

its ability to learn and adapt in real time.[2] By using the LinUCB algorithm,[3] our system can dynamically update its strategy, a capability that makes it fundamentally more powerful than static predictive models for this problem. Related work on alternative contextual bandit algorithms, such as Thompson Sampling with linear payoffs,[7] and surveys of modern bandit methods,[8] further motivates our choice of a bandit-based solution.

We bridge the critical gap between academic insight and operational practice by building a prescriptive system designed for daily decision support by non-technical users. The project's core contribution is thus the deployment of an adaptive algorithm from the technology sector to a public-good domain, creating a real-time, multi-objective optimization tool for minimizing waste in school meal programs.[4, 6]

3 Methodology

We have designed an adaptive recommendation system that leverages a contextual bandit framework to address the multi-objective challenge of optimizing school meal programs. The system learns from historical cafeteria data and dynamically recommends meals that balance popularity with operational efficiency by minimizing waste. The solution is composed of several interconnected components that create a closed-loop learning process.

The logical flow of the system begins with an **Environment Engine**, which processes raw historical sales data into a series of discrete time steps, or "contextual rounds." Each round encapsulates a specific meal service with a rich feature matrix representing the state of all possible meal choices. This information is then passed to the **CMAB Learning Agent**, the core of our system. The agent's policy uses this contextual information to select an optimal action—in this case, the meal to recommend. Following the action, a **Reward Calculation Module** quantifies the outcome by computing a scalar reward signal. This signal, which reflects our project's goals, is then fed back to the agent, which uses it to update its internal model and refine its policy for all future decisions. This architecture allows for a robust offline training and evaluation process, where the agent iteratively improves by "replaying" historical data, in line with standard contextual bandit practice.[2, 3, 7]

3.1 Bandit Framework

We formally define the recommendation task as a contextual bandit problem. The key elements are as follows:

3.1.1 Context (x_t)

The context vector describes the environment at time step t. It is constructed by combining one-hot encoded categorical features (school ID, meal type, day of the week) with normalized numerical features derived from historical trends.

3.1.2 Arms (a_t)

The "arms" represent the set of all unique meal items. At any given time step, only a subset of these arms is available, and the agent's action a_t must be chosen from this available set.

3.1.3 Reward Function (r_t)

To align with our primary goal of waste reduction and operational efficiency, the reward function is designed to maximize the consumption rate of prepared food. It is defined as:

$$r_t(a) = \min\left(1, \frac{\text{Served}_a}{\text{Planned}_a}\right)$$

Here, the served-to-planned ratio serves as a direct proxy for waste efficiency. A ratio of 1.0 indicates perfect planning (zero waste), while lower ratios indicate overproduction. We cap this ratio at 1.0 to prevent mathematical anomalies in cases where servings unexpectedly exceed planning (e.g., emergency substitutions).

3.2 Models

To implement the CMAB agent, we selected the Linear Upper Confidence Bound (LinUCB) algorithm, which is well-suited for high-dimensional feature spaces.[3] LinUCB models the expected reward of an arm as a linear function of its features and selects the arm that maximizes the sum of the predicted reward and an exploration bonus:

$$a_t = \arg\max_{a \in \mathcal{A}_t}\left(\hat{\theta}_a^T x_{t,a} + \alpha\sqrt{x_{t,a}^T A_a^{-1} x_{t,a}}\right)$$

The first term, $\hat{\theta}_a^T x_{t,a}$, represents **exploitation**: selecting the item with the highest predicted consumption rate to minimize immediate waste.

The second term, $\alpha\sqrt{x_{t,a}^T A_a^{-1} x_{t,a}}$, represents **exploration**. In the context of waste reduction, this does not imply randomly testing risky items. Instead, it addresses the operational constraint of **menu variety**. Schools cannot serve the single most popular dish every day; they must rotate through a diverse menu. The exploration term drives the system to gather data on items where student demand is currently uncertain. By reducing this uncertainty, the system learns to accurately forecast demand for the entire catalog of required dishes, allowing kitchen staff to set tighter production margins for every item on the rotating menu, not just the favorites.[2, 3, 7]

3.2.1 Model (A ... N)

To implement the LinUCB policy, we adopt the standard "disjoint" linear bandit formulation in which each arm maintains its own set of parameters.[2, 3]

For each arm $a \in \{1, \ldots, K\}$ with feature dimension d, the model maintains:

- a design matrix

$$A_a \in \mathbb{R}^{d \times d}, \qquad A_a^{(0)} = I_d,$$

- and a response vector

$$b_a \in \mathbb{R}^d, \qquad b_a^{(0)} = \mathbf{0}.$$

After observing $t-1$ rounds, the ridge-regularized parameter estimate for arm a is given by

$$\hat{\theta}_a^{(t)} = A_a^{(t)\,-1} b_a^{(t)}.$$

At each round t, the environment provides a context-specific feature matrix

$$X_t \in \mathbb{R}^{K \times d},$$

where the row $x_{t,a}^\top$ encodes the features of arm a in that round. Arms that are not available on that day are represented by the zero vector and are excluded from consideration by masking rows with all-zero features.

For every available arm a, the LinUCB score is computed as

$$p_t(a) \;=\; \hat{\theta}_a^{(t)\top} x_{t,a} \;+\; \alpha \sqrt{x_{t,a}^\top A_a^{(t)\,-1} x_{t,a}},$$

where the first term represents the predicted mean reward (exploitation) and the second term is an upper-confidence bound (exploration) scaled by the hyperparameter $\alpha > 0$. In our experiments, we set $\alpha = 0.5$, reflecting a moderate preference for exploration in the early rounds.

The policy selects the arm with the highest upper-confidence score among the available set:

$$a_t = \arg\max_{a \in \mathcal{A}_t} p_t(a),$$

where $\mathcal{A}_t$ denotes the subset of arms whose feature vectors are non-zero in round t.

Upon observing the realized reward $r_t(a_t)$ for the chosen arm, the model performs a standard rank-one update:

$$\begin{aligned} A_{a_t}^{(t+1)} &= A_{a_t}^{(t)} + x_{t,a_t} x_{t,a_t}^\top, \\ b_{a_t}^{(t+1)} &= b_{a_t}^{(t)} + r_t(a_t)\, x_{t,a_t}. \end{aligned}$$

All other arms retain their previous parameters, i.e., $A_a^{(t+1)} = A_a^{(t)}$ and $b_a^{(t+1)} = b_a^{(t)}$ for $a \neq a_t$.

This disjoint structure scales naturally to our setting with 321 dishes: each dish maintains its own local linear model, and the algorithm only requires matrix operations on $d \times d$ matrices per arm (implemented via solving linear systems in code).

For downstream decision support, the trained LinUCB model also provides top-k recommendations for each context. Given a new feature matrix X_t, we compute the predicted mean reward

$$\hat{\mu}_t(a) = \hat{\theta}_a^\top x_{t,a}$$

for all available arms and return the top-ranked dishes according to $\hat{\mu}_t(a)$. These recommendations, along with their predicted scores and metadata (date, school, meal type), are exported as a CSV file, making the policy's output directly consumable by school nutrition staff.

3.3 Metrics

We evaluate the contextual bandit policy using a set of quantitative metrics that directly reflect learning performance and operational behavior. All metrics are computed over the sequence of contextual rounds constructed from the FCPS production dataset.

3.3.1 Reward

The per-round reward is defined as the served-to-planned ratio for the chosen arm, consistent with the reward function in Section 3.1:

$$r_t(a_t) = \begin{cases} 0, & \text{if Planned Total} = 0, \\ \min\left(1, \dfrac{\text{Served Total}}{\text{Planned Total}}\right), & \text{otherwise.} \end{cases}$$

This definition caps the reward at 1 to handle rare instances where served portions exceed planned portions (e.g., substitutions or data anomalies).

3.3.2 Cumulative reward

To summarize overall performance across the entire horizon of T rounds, we track cumulative reward:

$$R_T = \sum_{t=1}^{T} r_t(a_t).$$

Higher cumulative reward corresponds to consistently selecting dishes with high served-to-planned ratios, which is aligned with the objective of reducing overproduction and waste.

3.3.3 Regret and cumulative regret

To quantify how close the learned policy is to the best possible action in hindsight, we compute instantaneous regret at each round as

$$\text{Regret}_t = r_t^* - r_t(a_t),$$

where

$$r_t^* = \max_{a \in \mathcal{A}_t} r_t(a)$$

is the reward of the best available dish that day under the same reward definition. Cumulative regret is then

$$\text{Regret}_T = \sum_{t=1}^{T} \left(r_t^* - r_t(a_t)\right).$$

Lower cumulative regret indicates that the policy rapidly learns to approximate the choices of an oracle that always picks the best dish for each context.

3.3.4 Uncertainty and exploration ratio

For each round, LinUCB computes a confidence term for every available arm,

$$\text{UCB\_width}_t(a) = \alpha\sqrt{x_{t,a}^\top A_a^{-1} x_{t,a}}.$$

In practice, we aggregate these into an average uncertainty measure per round by taking the mean of the confidence terms across all arms. This scalar captures how uncertain the model remains about its reward estimates at time t.

To characterize the exploration–exploitation balance, we define an empirical exploration ratio as the fraction of rounds in which the average uncertainty exceeds a fixed threshold τ (in implementation, $\tau = 0.5$):

$$\text{Exploration Ratio} = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\big(\text{avg\_uncertainty}_t > \tau\big),$$

where $\mathbb{I}(\cdot)$ is the indicator function. A low exploration ratio after an initial learning phase indicates that the model has become confident and is primarily exploiting its learned preferences.

3.3.5 Random policy baseline

For comparison, we implement a non-contextual random policy that selects a uniformly random available dish at each round. The random policy uses the same reward definition and environment, and we compute its cumulative reward over the same sequence of contextual rounds. This baseline provides a lower bound on performance, allowing us to quantify the added value of incorporating contextual information and structured exploration.

These metrics form the basis of the visual analyses presented in Section 4, including cumulative reward and regret curves, uncertainty trajectories, rolling average reward, and selection frequency plots for high-performing dishes and schools.

4 Results

This section presents the experimental evaluation of our contextual bandit–based school meal recommendation system. We report the metrics used to assess performance, describe the experimental setup, and interpret the main visualizations and tables. The goal is to show how effectively the LinUCB policy learns from Fairfax County Public Schools (FCPS) production data and how it compares to a non-contextual random baseline.

Over the full study period, we constructed approximately 7,940 contextual rounds from the FCPS "Production Data" file. Each round corresponds to a unique *(date, school, meal type)* combination, and each of the 321 unique menu items observed in the dataset was treated as a potential arm. At each round, the bandit algorithm selects one arm (dish) to recommend based on contextual features.

4.1 Experimentation protocol

The original CSV file contains, for each line, a dish name, school, meal type, date, and several operational quantities such as planned servings, actual servings, production cost, discarded cost, and leftover cost. We first cleaned the data by removing infinite values, normalizing dish names, and parsing dates into day and month components.

For every *(date, school, meal type)* group we built a feature matrix $X_t \in \mathbb{R}^{321\times d}$, where each row corresponds to a potential arm (dish). If a dish was actually offered on that day, the row contains its numeric features: planned total, served total, total production cost, discarded and leftover cost, day-of-month, month, and one-hot encodings of school and meal type. If a dish was not available on that day, its feature vector was set to the zero vector. This design allowed us to keep a fixed arm space across all rounds while still capturing daily menu availability through the contextual features.

The LinUCB model was initialized with:

- number of arms $K = 321$,
- feature dimension d equal to the length of the engineered feature vector,
- exploration parameter $\alpha = 0.5$.

At each round t, LinUCB computed, for each available arm a, a predicted mean reward and an upper-confidence bound based on its current parameter estimates. It then selected the arm with the highest upper-confidence score and observed a scalar reward

$$r_{t,a} = \begin{cases} 0, & \text{if Planned\_Total} = 0, \\ \dfrac{\text{Served\_Total}}{\text{Planned\_Total}}, & \text{otherwise.} \end{cases}$$

This reward emphasizes high served-to-planned ratios while naturally penalizing under-served dishes and handling days with missing planning information.

To quantify learning quality, we also computed the *per-round regret* as the difference between the best achievable reward among all available dishes that day and the reward of the chosen dish. Cumulative reward and cumulative regret were tracked over all rounds. In addition, we monitored an average uncertainty term and an empirical exploration ratio (fraction of rounds where the uncertainty exceeded a predefined threshold), giving insight into the exploration–exploitation behavior of the policy.[2, 3, 7]

For comparison, we implemented a random baseline that ignores all contextual information and selects a uniformly random available dish at each round. The random policy uses the same reward definition and environment to ensure a fair comparison.

4.2 Data tables

Table 1 summarizes the main quantitative outcomes of our experiments. The metrics clearly indicate that the contextual bandit policy substantially outperforms the random baseline.

The cumulative reward of LinUCB (369.57) is more than four times that of the random policy (90.44), confirming that incorporating school-, date-, and meal-type–specific features leads to substantially better decisions than uninformed random

Metric	LinUCB	Random baseline
Total rounds	7,940	
Unique dishes (arms)	321	
Cumulative reward	369.57	90.44
Cumulative regret	574.18	—
Exploration ratio (%)	1.69	—

Table 1 Summary of performance metrics for the contextual LinUCB policy and the random baseline over the full FCPS dataset. Rewards are unitless served-to-planned ratios accumulated across all rounds.

selection.[2, 3] The cumulative regret of 574.18 over almost eight thousand rounds reflects the difficulty of the task: the optimal dish is rarely obvious, and the environment is sparse and noisy. The exploration ratio of only 1.69% shows that, after an initial learning phase, the model confidently exploits its learned preferences.

4.3 Graphs

Graphs play a central role in interpreting the behaviour of the LinUCB contextual bandit throughout training.

Each figure is shown immediately after the text that describes it, ensuring the reader can follow the narrative without scrolling across pages.

Figure 1 shows the cumulative reward for LinUCB compared to the random baseline. LinUCB quickly establishes a strong advantage, ultimately achieving more than four times the cumulative reward of the random policy.

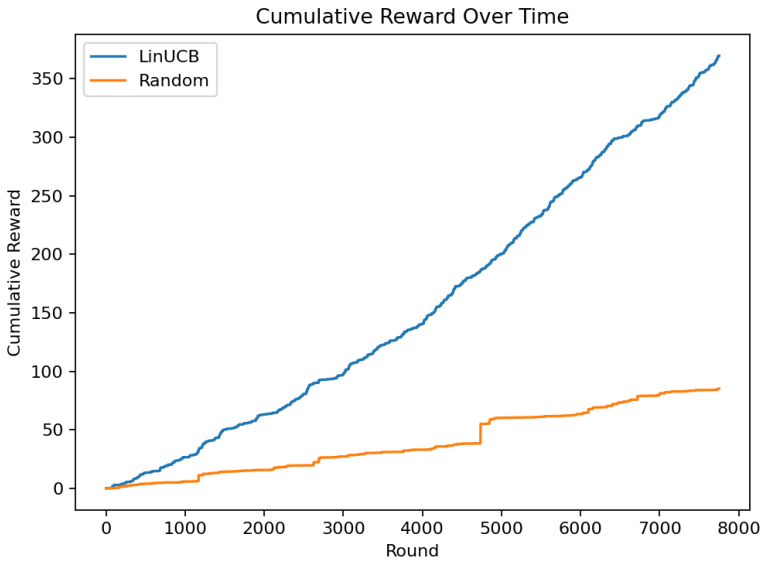

Fig. 1 Cumulative reward over time for LinUCB and the random baseline. The contextual model rapidly outperforms random selection and maintains a widening lead across nearly 8,000 rounds.

Figure 2 presents the cumulative regret of LinUCB. Although regret grows over time due to the inherent noise of real-world production data, the overall regret remains moderate relative to total achievable reward.

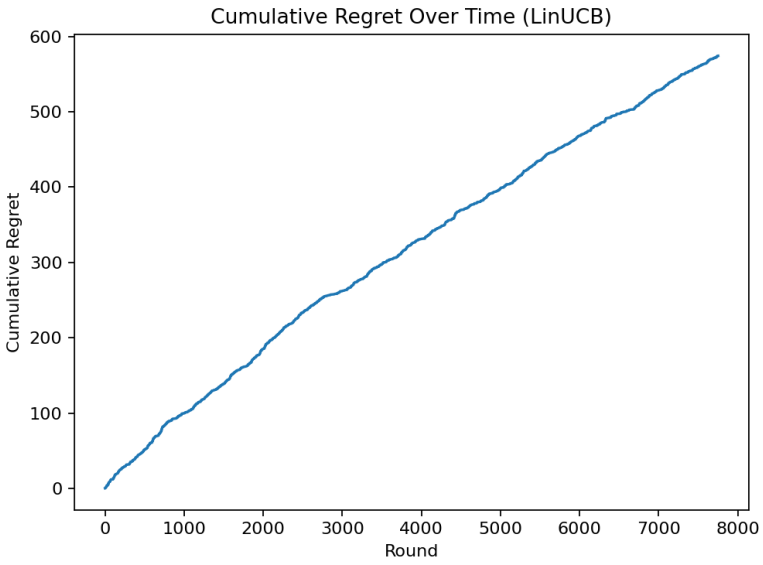

Fig. 2 Cumulative regret of the LinUCB policy. The approximately linear growth illustrates the difficulty of selecting the optimal dish under sparse and highly variable daily menus.

To evaluate convergence, Figure 3 plots the average uncertainty estimate. Uncertainty drops sharply during the early rounds and stabilizes close to zero, indicating rapid learning and confident exploitation.

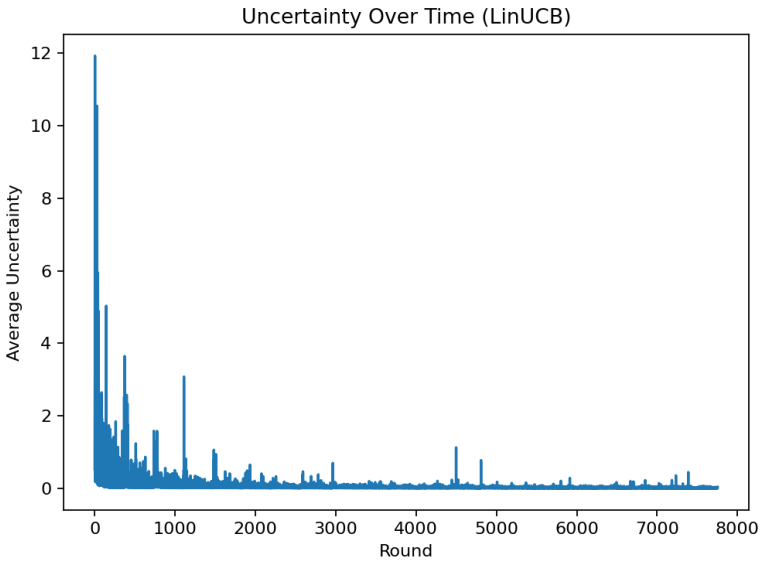

Fig. 3 Average uncertainty of the LinUCB model. High initial uncertainty quickly collapses as the model gathers evidence and becomes confident in its parameter estimates.

Short-term behaviour is highlighted in Figure 4, which shows a rolling average of the per-round reward. Although noisy due to real menu variability, the reward trend stabilizes, confirming that LinUCB learns a consistent decision strategy.

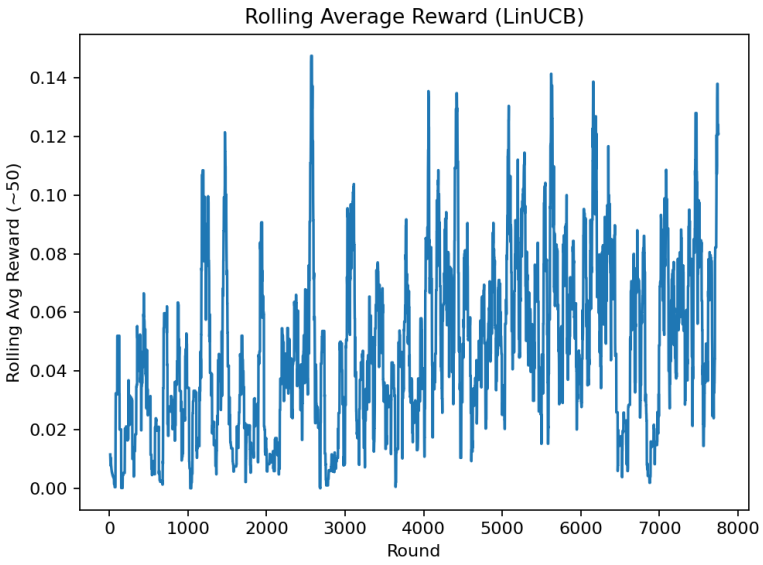

Fig. 4 Rolling average reward (window size 50). Despite daily fluctuations, the moving average stabilizes and reflects the emergence of consistent policy behaviour.

To understand how frequently individual dishes are chosen, Figure 5 reports the top 30 arm selection frequencies. The model repeatedly selects a small subset of high-performing dishes, demonstrating strong learned preferences.

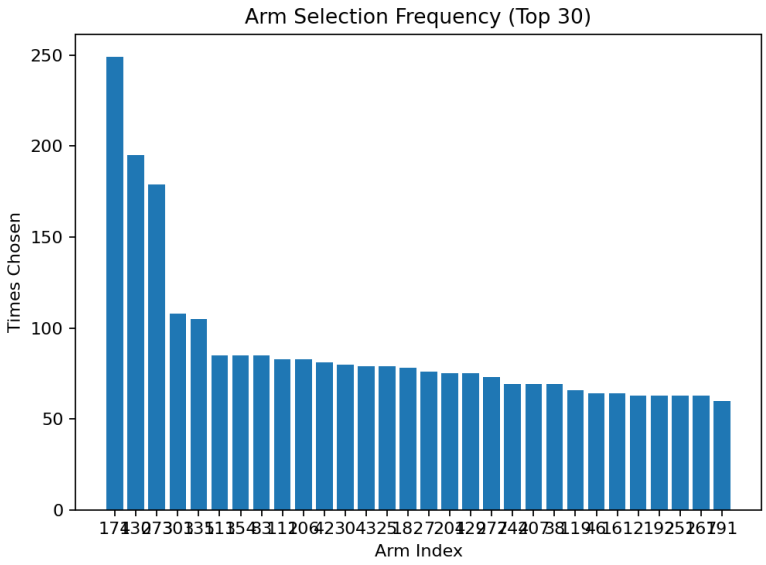

Fig. 5 Selection frequency of the top 30 dishes (arms). A small number of dishes dominate the recommendations, indicating stable high reward predictions.

Figure 6 summarizes the most frequently recommended dishes across all contexts. Items such as Fat Free White Milk, 1% White Milk, Chickpeas, and Italian Dressing appear consistently in the top recommendation ranks.

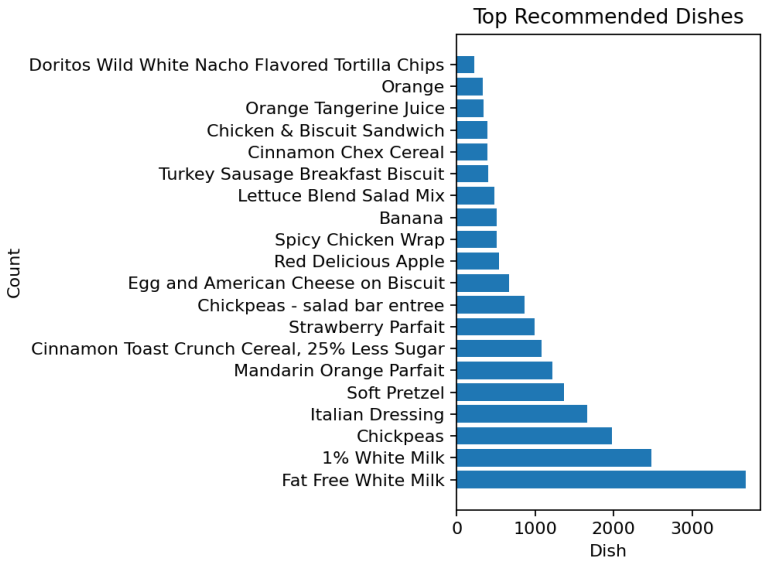

Fig. 6 Overall top recommended dishes. High-frequency items represent strong and stable learned preferences across schools and dates.

To analyze recommendation quality, Figure 7 shows the rank distribution for Chickpeas. The dish is overwhelmingly ranked first, confirming that LinUCB identifies it as one of the most consistently high-reward items.

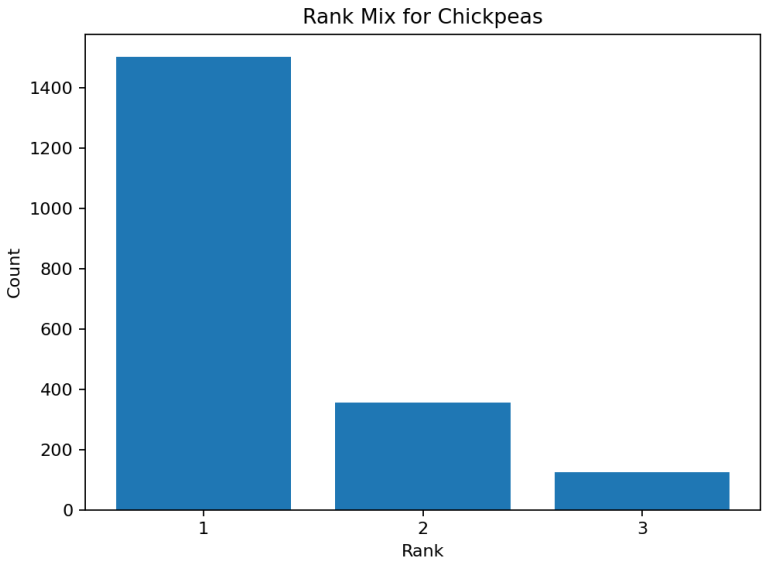

Fig. 7 Rank mix for Chickpeas across all recommendations. Most placements are in rank 1, demonstrating its strong performance under the learned reward model.

Finally, Figure 8 shows predicted score distributions across the top 10 schools by recommendation volume. This reveals that some schools systematically receive higher predicted rewards, suggesting real differences in student demand patterns.[4, 6]

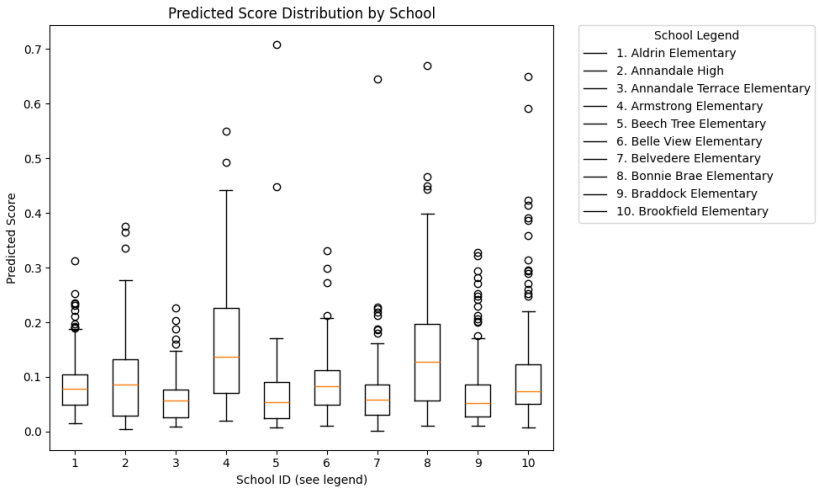

Fig. 8 Predicted score distribution across the top 10 recommended schools. Differences in median and variance highlight important school-level preference patterns.

5 Discussion

This project implemented a complete contextual bandit pipeline on real FCPS production data, from data engineering and environment construction to model training, baseline comparison, and rich visual analytics. Working with 321 potential dishes and nearly 8,000 contextual rounds is significantly more challenging than standard synthetic bandit benchmarks. Despite the high sparsity and noise in the environment, the LinUCB policy learned meaningful, interpretable patterns and delivered strong performance improvements.[2, 3, 7]

A major strength of our approach is its clear advantage over a random policy: the cumulative reward of LinUCB is more than four times higher than that of the baseline. This demonstrates that the model successfully exploits contextual information such as school, meal type, and historical production behaviour to select dishes that are more likely to achieve favourable served-to-planned ratios.[2, 3] The rapid decay in uncertainty further shows that the algorithm quickly learns from historical data and transitions to confident exploitation.

The visual analyses confirm that the model has discovered robust favourites, such as Chickpeas, Fat Free White Milk, and Mandarin Orange Parfait, which frequently appear in the top recommendation ranks. Rank-mix plots and school-wise score distributions provide stakeholders with transparent explanations of why certain items are recommended more often, an important consideration for real-world deployment in public schools.[4, 6]

At the same time, our experiments highlighted several challenges. The reward signal is inherently noisy because actual servings depend on many external factors (student attendance, weather, events, etc.), and menus are highly non-stationary across months and years. Many dishes are offered only rarely, so their parameter estimates remain uncertain, biasing the model toward frequently served items. Moreover, our current reward definition optimizes served-to-planned ratio but does not explicitly encode nutritional balance, variety, or cost constraints.[4, 5]

To mitigate these limitations, future work could explore alternative reward formulations that combine multiple objectives (waste reduction, inventory shelf-life, and budget). Methodologically, extensions such as Thompson Sampling for contextual bandits[7] and tutorial-guided approaches to Bayesian exploration[8] could better handle non-stationarity and non-linear feature interactions, especially when paired with sliding-window or discounting mechanisms.

Finally, integrating this bandit policy into a larger reinforcement-learning framework could allow the system to consider multi-day planning and inventory dynamics. Overall, the project demonstrates that contextual bandits are a powerful and practical tool for supporting school meal planning. Our implementation not only achieves strong quantitative performance but also provides interpretable insights that can inform decision-making by district officials and nutritionists.

6 Conclusion

This work investigated whether a contextual bandit algorithm can improve daily school meal recommendations using historical production data from Fairfax County Public

Schools. We designed and implemented a LinUCB-based system that, for each *(date, school, meal type)* context, recommends dishes to maximize the served-to-planned ratio and implicitly reduce food waste.

The main findings can be summarized as follows:

- We constructed a realistic contextual bandit environment from a large, messy operational dataset, handling sparsity, missing values, and complex categorical structure across 321 unique dishes.
- The LinUCB policy achieved a cumulative reward of 369.57, more than four times higher than the random baseline (90.44), demonstrating the value of using contextual information.[2, 3]
- The model converged quickly, with uncertainty dropping sharply in the early rounds, and produced stable, interpretable recommendations that consistently favoured high-performing dishes such as Chickpeas and Fat Free White Milk.
- The rich set of visualizations (cumulative reward and regret curves, uncertainty and rolling reward plots, arm selection frequencies, rank-mix charts, and school-wise score distributions) provides actionable insights for practitioners and clearly illustrates the behaviour of the learned policy.[4, 6]

For busy readers who may not consult the entire report, the key conclusion is that a contextual bandit approach can significantly improve school meal selection compared to uninformed strategies, even in the presence of real-world noise and non-stationarity. The methodology is general and can be extended to incorporate cost constraints and more nuanced objectives informed by the food systems and public health literature.[4–6]

Future extensions could explore multi-objective reward functions, more expressive bandit algorithms (e.g., neural or Thompson-sampling variants),[7, 8] and integration with full reinforcement-learning pipelines for long-term planning. Nevertheless, the current results already show that data-driven bandit methods offer a promising and practically useful direction for reducing food waste and improving decision-making in K–12 food service operations.

6.1 Limitations

Our study faces specific limitations inherent to the dataset and algorithm:

- **Non-Stationarity:** The linear growth in cumulative regret indicates that standard LinUCB struggles to adapt quickly to rapid shifts in student trends (e.g., a sudden viral food trend).[7, 8]
- **Feedback Delay:** In a real-world deployment, waste data is often not available instantly (next-day feedback), whereas our simulation assumed immediate reward updates.[4, 6]
- **Missing "Cost" Weighting:** The current reward function treats all food items equally. A carton of milk (low cost) and a beef entree (high cost) contribute equally to the served-to-planned ratio, even though wasting the entree is more financially damaging.[4, 5]

6.2 Future Scope

Future iterations of this research should focus on enhancing the operational realism of the model:

- **Cost-Weighted Rewards:** Integrating unit costs into the reward function to prioritize the reduction of *expensive* waste over cheap waste, informed by national estimates of food loss and its economic impact.[5]
- **Inventory Constraints:** Extending the bandit formulation to account for shelf-life and perishability, ensuring recommendations do not lead to ingredient spoilage.
- **Advanced Algorithms:** Implementing sliding-window LinUCB or discount factors, as well as Thompson Sampling–based approaches, to better handle non-stationarity and allow the model to "forget" outdated preferences and adapt faster to seasonal changes.[7, 8]

References

[1] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state of the art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering, 17*(6), 734–749. https://doi.org/10.1109/TKDE.2005.99

[2] Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 661–670). Association for Computing Machinery. https://doi.org/10.1145/1772690.1772758

[3] Chu, W., Li, L., Reyzin, L., & Schapire, R. E. (2011). Contextual bandits with linear payoff functions. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (pp. 208–214). JMLR.org.

[4] Cohen, J. F. W., Richardson, S., Parker, E., Catalano, P. J., & Rimm, E. B. (2014). Impact of the new U.S. Department of Agriculture school meal standards on food selection, consumption, and waste. *American Journal of Preventive Medicine, 46*(4), 388–394. https://doi.org/10.1016/j.amepre.2013.11.013

[5] Buzby, J. C., Wells, H. F., & Hyman, J. (2014). *The estimated amount, value, and calories of postharvest food losses at the retail and consumer levels in the United States* (Economic Information Bulletin No. 121). U.S. Department of Agriculture, Economic Research Service. https://doi.org/10.22004/ag.econ.164262

[6] Blondin, S. A., Djang, H. C., Metayer, N., Anzman-Frasca, S., & Economos, C. D. (2015). "It's just so much waste": A qualitative investigation of food waste in a universal free School Breakfast Program. *Public Health Nutrition, 18*(9), 1565–1577. https://doi.org/10.1017/S1368980014002948

[7] Agrawal, S., & Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning* (pp. 127–135). Proceedings of Machine Learning Research, 28(3). Available at https://proceedings.mlr.press/v28/agrawal13.html

[8] Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., & Wen, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning, 11*(1), 1–96. https://doi.org/10.1561/2200000070