**Sam’s Club – Attribute Based Assortment**

Noa Scharf, Hector Negron, Benjamin Tutka

University of Arkansas

DASC 49903 – DASC Practicum II

Dr. Schubert

05/07/2025

**Abstract**

Managing space and product assortment across 596 Sam’s Club locations presents a significant challenge due to the diverse nature of the club’s formats, driven by their historical factors such as acquisitions and strategic shifts. Despite Sam’s Club recent growth in in-club revenue (48%) and e-commerce sales (150%) since FY19, the space allocation strategy has largely relied on “top-down” models. However, with changing market dynamics and member expectations, there is a need to evolve this strategy. To do so, this project will explore a bottom-up approach to space and assortment planning, by leveraging item-level attribution data and member demographics.

Using machine learning, the goal is to develop assortments for Sam’s Clubs, optimizing space allocation based on member profiles and aligning product assortments with financial performance. Key questions include understanding how similar assortments and layouts affect demographic alignment, measuring the impact of this alignment on sales, and identifying discrepancies in clubs where product attributes don’t meet member needs.

**Background and Information**

Sam’s Club is a membership-based warehouse club, which was founded by Sam Walton with the vision of delivering value and quality to its members. The business focuses on providing a streamlined shopping experience through a curated assortment of products, allowing members to access significant savings (in bulk). Today, Sam’s Club operates nearly 600 locations in the United States and Puerto Rico, as well as in China and Mexico, serving millions of members both in-store and online. The company has seen significant growth in recent years, as mentioned above. This growth, alongside a variety of unique store layouts and a shift toward multi-channel fulfillment options, has created challenges in optimizing space and product assortment. To address these complexities, Sam’s Club is exploring a shift from a traditional top-down space allocation model to a data-driven, “bottom-up” approach. This involves leveraging item-level data and machine learning to create tailored assortments that better align with the demographics and preferences of each location. Through this approach, Sam’s Club aims to enhance the member’s experience and improve financial performance by aligning space and assortment with local needs.

From our project proposal and meetings with our Sam’s Club team, some of the key questions we are asking include:

1. Are certain clubs aligning their assortment more effectively with their member profiles?
   1. Does this alignment correlate with financial performance?
2. Do there exist relationships between club attributes and certain subcategory sales?
   1. If so, are the relationships common for all subcategories, or variable across GMMs?

**Previous Approaches Research**

McKinsey's approach to assortment optimization focuses on using advanced analytics to maximize retail efficiency by aligning product assortments with customer demand and store-specific needs. They use tools like customer decision trees and a proprietary "walk rate" metric to understand how removing a product might impact sales, helping to balance inventory and avoid overloading shelves with low-demand items. Their system includes evaluating factors such as financial contribution, product uniqueness, and cost-to-serve, enabling retailers to make informed decisions on which products to list, remove, or expand. This method not only boosts profit margins and sales but also simplifies supply chain processes by reducing SKU complexity. Through pilot programs, McKinsey demonstrates quick wins that make this analytical approach both feasible and valuable, paving the way for retailers to use data-driven insights as an integral part of their ongoing assortment strategies (McKinsey)​.

Kroger’s Assortment & Space Recommender (KASpR) leverages machine learning to enhance store layouts by making merchandising decisions that align with evolving customer preferences. Traditional solutions fell short because they lacked item-level data, limiting Kroger’s ability to understand assortment-specific performance and quantify the financial impact of layout changes. By applying its proprietary machine learning approach, KASpR predicts how different planogram options will affect category sales in specific stores. This system factors in unique store attributes, like fixtures and space constraints, to create tailored recommendations. KASpR’s data-driven method reportedly contributes an estimated $18 million in annual sales through layout optimization in remodels and other projects. By offering store-specific sales predictions, KASpR empowers Kroger to use data insights across various merchandising projects, ensuring decisions are both strategic and impactful (*Optimizing store space with machine learning*).

**Understand**

The goal of our project is to make recommendations for assortment changes at the item-level. These changes could be as simple as changing the position within the assortment (moving an item up or down), or as extreme as removing an item from certain stores. The data necessary for this analysis will be the sales data from items of interest (designated by Sam’s Club), and demographic data pertaining to the individual Sam’s Clubs. One potential solution will be to calculate the average sales of an item across all Sam’s Clubs (scaled by total sales of the club), and then identify outliers above or below the average sales threshold. Further investigation of those clubs may help us understand why certain items are over or underperforming in particular stores. Success for this project will look like item-level assortment recommendations based on machine learning models that will help Sam’s Club increase sales for some items and remove items that are underperforming.

**The Data**

The data that we will be using in this project includes many types, including Sam’s Club internal data (about member demographics, club attributes, club competitor’s attributes) as well as feature engineering data that was personally found by team members. The feature engineering data includes US demographic data, economic data, and employment data. Later on in the project, we gathered NCAA stadium data from Kaggle, to create a variable which would serve as a potential factor influencing sales performance.

The sources of our data vary as well, for example, the Sam’s Club specific data came from the internal team at Sam’s Club that our team is working with. For the feature engineering data, we found our data from the US Census Bureau and Federal Reserve Economic Database (FRED).

**Figure 1.**

*A screenshot of a computer

Description automatically generatedSample photo of the merged data (club data, and club physical attribute data)*

As seen above, the club data and the attribute data about the clubs were merged in a Jupyter notebook, and output into a single csv. A snippet of the output csv is seen in [Figure 1](#Figure1). Our data includes variables about store-level sales (item, subcategory, and total sales), as well as store attributes like club names, region numbers, state, urbanicity, market area, and climate zones. The additional demographic and competitive variables that were included in our data were population, total median earnings, household growth, store class, and competition proximity (to Costco and Walmart). The data also included a few geospatial variables such as latitude and longitude, proximity to Power 5 stadiums, military bases, and coastal or lake areas.

Because this study integrated data from multiple sources, extensive preprocessing was necessary to ensure consistency and reliability. The initial step addressed missing values by removing rows with null entries to preserve data accuracy. Next, item-level and store-level data from Sam’s Club were merged, while Census and stadium datasets were integrated based on geographic proximity. To enable comparisons across store locations, sales data was aggregated at the store level, ensuring that both item-level and subcategory-level contributions were accurately captured.

Feature engineering was essential for enhancing the dataset. A new variable was introduced to indicate each store’s distance from a Power 5 stadium, calculated using geospatial distance metrics. Additional features were created to represent item and subcategory sales as a percentage of total store sales, helping normalize performance comparisons across high- and low-volume stores and reduce noise in the models. Clustering techniques were then used to group stores with similar demographic and competitive profiles.

To prepare the data for modeling, multi-fold cross-validation was employed for feature selection, which helped identify the most impactful predictors while minimizing the risk of overfitting. Throughout the preprocessing process, careful attention was paid to avoid duplicating sales data when merging datasets. Since sales information was available at both the item and store levels, the aggregation process was managed to prevent any distortion in total sales figures.

**Semester 1 Results**

For semester one, we wanted to start with some results from our exploratory data analyses. We were given 2 categories to focus on, a fresh category (specifically floral and produce), and a general merchandise category (specifically seasonal décor).

**Figure 2.**

*Masked EDA graph of floral and produce sales over a 2-year range (2022-2024)*

A blue line on a white background

Description automatically generated

From [Figure 2](#Figure2), you can see some peaks that follow the same patterns over the two years. These peaks are from Mother’s Day, and Valentine’s Day. As seen in the visual, the low points are days that Sam’s Clubs were closed. We also did a deep dive into the subcategories sales, but felt that they were not as useful to us at this point, because we would like to group them by clusters, which will be happening early next semester, right before modeling.

**Figure 3.**

*Masked EDA graph of seasonal décor sales over a 2-year range (2022-2024)*

A graph showing a graph

Description automatically generated with medium confidence

As seen above, [Figure 3](#Figure3) covers seasonal décor’s sales over the years of 2022 – 2024, for all the Sam’s Clubs. In the graph, we see peaks during certain points of the years, with the rest of the year being low points. This is because of the category (seasonal décor), the sales begin to spike during August, peaking in December, and then going back down starting mid-January. These peaks repeat over the 2-year range, hinting at a trend (which was confirmed by our industry partner).

The two figures above were the most insightful results that we received over our analyses. We performed clustering analyses on the 600 clubs, using a multitude of attributes, to create 7 clusters to mimic the 7 market areas that already exist.

**Figure 4.**

*Clustering analysis for 600 Sam’s Club stores*

A map of the united states with colored dots

Description automatically generated

We would like to use these clusters seen in [Figure 4](#Figure4) to go deeper into our subcategory analyses, seeing which ones perform best in the different clusters.

**Semester 1 Retrospectives**

Now for our key takeaways and retrospectives for the semester, one of our biggest insights was that plans may change along the way. Towards the beginning of our project, we were focused on a different dataset for our POS sales, which we found out has not been recently updated. Therefore, we had to change paths and familiarize ourselves with a new dataset covering Sam’s Club sales. Some of our other key takeaways included the trends we found within the two categories we dove into (mentioned above) and finding leading items for subcategories across all stores, which is changing to be grouped by clusters.

What we think went well this semester was our planning, and the exploratory data analyses that we ended up with. Our feature engineering efforts, data collection and combination, and analysis of the data is something that we are very happy with for the first semester. It has set us up for a smooth entry into our modeling portion, happening next semester. Though there was a rocky start, having to share code amongst each other (no access to Databricks), once access to Databricks was given by Sam’s Club, everything ran much smoother. Another thing that we feel went very well this semester would be our Trello board organization, this helped tremendously for our project management.

As for what we would have done differently this semester, asking for help earlier with some of the roadblocks we hit, specifically with data queries to grab the required data would be something that we would have done different. We also wish we would have clarified the end goals and primary objectives of each sprint, instead of simply assigning tasks for the sprints. Though on some sprints we had milestones, we wish we would have done that on every sprint in our Trello board. Adding onto our sprint clarification, we wish we specified the role that each sprint plays into fulfilling our end goal, which plays along with specifying end goals and objectives for sprints.

In the future, there are a few items that we want to get done, the first item being adding the clusters we created onto the end of our sales data. This will allow us to take a deeper dive into the categories we were assigned and create cluster-specific suggestions. Speaking with Michael, our industry partner, we also want to answer the question “why?”, we want to figure out why some items are outperforming in some clubs relative to other clubs, not just the simple fact that items outperform each other. If time allows, expanding our zip code to not only include the specific zip code that the club is in, but expanding it to the surrounding areas, since there are people who visit clubs from surrounding areas if their zip code doesn’t have a club. After this, our last item that we would want to complete (after modeling and drawing initial results) would be to expand the items that are analyzed.

**Model**

During our modeling phase, we have been tasked with exploring multiple machine learning models to predict item sales using our data. To ensure a comprehensive analysis, we employed both linear and non-linear approaches, including Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and CatBoost. Linear Regression served as a baseline model, capturing straightforward relationships between variables. The Decision Tree provides an interpretable yet non-linear approach, while the Random Forest takes a step further and enhances predictive performance by aggregating multiple trees. Gradient Boosting and CatBoost both refine predictions through iterative improvements, with CatBoost specifically optimized for handling categorical features efficiently (which there are fairly many in our dataset). Through comparative analysis, we aim to identify the most effective model(s) based on evaluation metrics.

The models in this study were designed to predict two target variables: the percentage of total store sales attributed to each item and to each subcategory. Rather than prioritizing predictive accuracy alone, the primary objective was to uncover which store-level attributes most significantly influenced item and subcategory sales. By comparing feature importance across models, the analysis aimed to identify key factors driving product performance.

To evaluate model performance, the dataset was split into 80% for training and 20% for testing. Standard metrics such as R² and mean squared error (MSE) were calculated to measure prediction accuracy, but the focus remained on interpreting feature importance. The goal was not to create the most accurate predictive model, but to understand which store characteristics had the strongest influence on sales distributions.

Modeling and data processing were carried out using Python, SQL, and PySpark. Python supported data analysis, machine learning, and visualization tasks, while SQL was used for querying and preparing the data. PySpark enabled scalable processing of large datasets. Databricks served as the cloud-based platform that integrated these tools, providing the infrastructure needed for efficient computation and model execution on a large scale.

Several challenges emerged during the research, particularly during the preprocessing phase. Merging multiple datasets required careful attention to avoid double-counting sales when aggregating from item- to store-level data. Feature engineering also posed difficulties, as it involved selecting variables that were relevant without introducing noise that could degrade model performance. Additionally, the size and complexity of the data demanded optimization techniques to manage computational load effectively.

Although no explicit assumptions were initially stated, the analysis relied on several implicit assumptions. First, it assumed that consumer purchasing patterns remained relatively consistent during the time period studied, allowing historical sales data to serve as a reliable indicator of future trends. It was also assumed that the available datasets were sufficiently complete to capture the key drivers of store sales. Lastly, the analysis presumed that geospatial proximity to Power 5 football stadiums could meaningfully affect consumer behavior, even though the strength of this effect likely varied by product category.

**Understand**

Individual models were developed for each subcategory using data from all Sam’s Club locations that generated over $100,000 in sales for that subcategory during 2024. After training, the coefficients for each model were saved as .pkl files for future analysis. These models support multiple levels of insight. At the broadest level, they reveal which features are most influential across the entire dataset—that is, across all subcategories and store locations. The most impactful factors across the full dataset included regional classification, surrounding population income, population density, an engineered cluster variable, the presence of nearby competitors, total population size in the area, and average seasonal snowfall. (Seen in [Figure 5](#Figure5), specific feature names have been masked.)

**Figure 5.**

*Top 10 Feature importances graph*

A graph with blue and white stripes

Description automatically generated

The analysis was then further refined at the General Merchandising Management (GMM) level, which groups related subcategories under broader classifications, such as Fresh or Seasonal Décor. Boxplots were used to visualize how the importance of individual features varied across GMMs. These visualizations allowed for direct side-by-side comparisons, highlighting differences in feature impact across merchandising groups. Boxplots also helped flag outliers—features with disproportionately high influence in certain subcategories.

These outliers were especially critical for deeper insight, as they showed where certain factors had an unusually strong effect. Examples of such high-impact features included average seasonal snowfall, total median earnings, regional placement, tropical climate zone, number of days open, household growth rate, hot climate zone, presence of nearby competitors, and proximity to military bases. (Seen in [Figure 6](#Figure6), again, specific feature names have been intentionally masked.)

**Figure 6.**

*Graphs comparing 4 different GMMs and their feature importances*

A screenshot of a graph

Description automatically generated

**Concluding Remarks**

This analysis serves as an initial step in helping Sam’s Club understand the complex relationships between store and customer attributes and item-level sales performance. By integrating multiple data sources and leveraging machine learning models, this study identified key drivers of product demand across subcategories and General Merchandising Management (GMM) groups. Although the models emphasized interpretability over raw predictive power, the insights they produced—such as the influence of regional demographics, competitor proximity, and climate factors—are both actionable and scalable. Substantial data preprocessing and validation were required, and future efforts should build on this foundation by refining the models, incorporating new variables, and expanding the scope of analysis. These preliminary findings validate the presence of meaningful patterns and provide a framework for ongoing assortment optimization. Sam’s Club plans to continue this initiative, using this project as a baseline for future student teams to enhance and extend the analysis in support of more precise, data-driven retail decisions.

**References**

Bureau, U. C. (2023, October 11). *Census datasets*. Census.gov. https://www.census.gov/data/datasets.html

McKinsey. (2019, October). <https://www.mckinsey.com/~/media/mckinsey/industries/retail/how> we help clients/big data and advanced analytics/mck\_retail\_analytics\_brochure\_v10.pdf

*Optimizing store space with machine learning*. Optimizing store space with machine learning | 84.51°. (n.d.). <https://www.8451.com/knowledge-hub/technology/optimizing-store-space-with-machine-learning/>

**Appendix**

A screenshot of a computer program

Description automatically generated

The processing function above was used to apply the CatBoost model to each subcategory across all Sam’s Club locations. It includes standard train-test splitting and the specification of categorical variables, which is a key requirement for CatBoost. Following model training, the resulting variable coefficients were saved as .pkl files using Python’s pickle module.

A screenshot of a computer program

Description automatically generated

This code adapted the Pandas-based processing function for use with Spark. Spark was chosen for several reasons: the Databricks Python driver frequently stalled during model training, Spark provided better handling of network interruptions, and its parallel processing capabilities significantly improved performance and efficiency.

A screenshot of a computer code

Description automatically generated

This code demonstrates how the .pkl files generated by the processing function were loaded for analysis. Some troubleshooting was required, as the code used to create the DataFrame initially produced errors. The issue was traced to the presence of an empty file named example.pkl, which contained no data and disrupted the loading process.

A screenshot of a computer

Description automatically generated

This code constructs the feature importances DataFrame by aggregating the feature coefficients from all models generated using the preprocessing function in Python and executed through Spark.