Optimizing Retail Strategy: Enhanced Product Association Modeling through Clustering and Market Basket Analysis for Dillard's

Authors: Daniel Wang, Mahi Shah, Tzuliang Huang, & Xinran Wang

Table of Contents

Executive Summary

In our project, we harnessed Dillard's comprehensive POS data, meticulously cleaning it and resolving inconsistencies to prepare for analysis. Our team clustered ~ 550,000 SKUs into three groups using K-prototype clustering, a method adept at managing both categorical and numerical data. We engineered 5 features: return_rate, size, department, most_sold_season and discounted to build our model. Then, applying Market Basket Analysis (MBA), we discovered significant association rules among products, with a support level of 0.089. Drawing insights from these association rules, we crafted targeted recommendations for customers who purchased specific products, encouraging them to consider associated items as well. This led to a discernible increase in Dillard's purchase rate from 3% to 3.1%. Our project not only demonstrates the value of precise data manipulation and advanced analytics in retail but also highlights the potential of data-driven strategies to incrementally boost business performance. Through our approach, Dillard's can now leverage these insights for more informed recommendation and marketing decisions (like what items can be discounted together), fostering a more engaging and customized shopping experience.

1. Introduction

In the dynamic landscape of retail, data-driven insights have become indispensable for businesses seeking to enhance their operational efficiency and customer satisfaction. We make an attempt to revolutionize Dillard's retail strategy by employing advanced data analytics techniques on its comprehensive Point of Sale (POS) data. Utilizing key datasets, such as SKUINFO for product attributes, STRINFO for store characteristics, TRNSACT for transactional histories, and SKSTINFO for inventory records, the company is optimizing product association rules. Through the application of clustering, we identify natural groupings of customer preferences and product performance, which enables more targeted recommendation strategies. Concurrently, Market Basket Analysis (MBA) deciphers patterns in product bundling and co-purchases. This dual analytic approach uncovers the association rules among different products, formulating targeted recommendations tailored to customers who had previously purchased specific products. By harnessing the richness of its POS data, we are not only enhancing the customer shopping experience through optimized recommendation strategies but also achieving a more efficient, data-driven business model that can dynamically respond to market demands.

2. Methods

The central framework of our model, illustrated in Figure 1, initiates a clustering process based on five carefully selected product features, such as color and department. Subsequently, for items grouped within the same cluster, we leverage Market Basket Analysis (MBA) using transaction history data to ascertain product association rules. Informed by these rules, our model provides targeted recommendations for basket composition and promotional strategies tailored to sellers' needs.

This two-step approach offers distinct advantages. Firstly, compared to applying MBA directly to the entire dataset, the preliminary clustering step significantly reduces computational efforts. By focusing on products within the same cluster, we streamline the association calculation process. Secondly, the integration of clustering enhances the precision of our recommendations. Notably, the practice addresses potential complexities in consumer behavior, such as adults purchasing items alongside their children. The application of clustering before MBA effectively mitigates such confounding factors, resulting in more accurate and actionable insights.

2.1 Data

The dataset includes store profiles (STRINFO), stock details (SKSTINFO), sales transactions (TRNSACT), product specifics (SKUINFO), and department classifications (DEPTINFO). See Figure 2 for ERD in appendix. We then cleaned the data, with major emphasis on the following tables:

- SKUINFO: Since there are some columns contained multiple values separated by commas, however the csv file we have also used comma to separate different columns, therefore, we used three unknown columns at the end to investigate on the rows that has data shifted to other columns using PACKSIZE as a reference point since it's easy to identify its values, shifted back the columns to the left for these affected rows. Investigate N/A and missing values inside the SKUINFO, replace with null values for easier access. After cleaning the data, we uploaded the data into the database.
- TRANSACT: Initially, it was observed that the "Sprice" and "AMT" columns are identical, leading to the decision to remove "Sprice". Additionally, columns with unspecified or unknown data were also eliminated to enhance data quality. A notable anomaly was identified in the "ORGPRICE" field, where a significant number of value were recorded as zero. This was an unusual finding, as logically, the original price of each SKU should not be zero. To address this discrepancy, these zero values were replaced with the corresponding SKU values, ensuring a more accurate and realistic representation of each SKU's original price. This adjustment not only rectifies the data irregularity but also aligns the dataset more closely with real-world retail pricing scenarios.

Prior to clustering the SKUs, 5 features were engineered to group similar SKUs together. The features were:

- Return Rate - Ratio of return of the item calculated using the TRANSACT data where returned items were flagged as those where type of transaction = R.
- Size - Numerical sizes from the SKUINFO table were bucketed into XS, S, M, L, XL, XXL & Other
- Department - Categorical variable from the DEPTINFO table divided into 3 departments. The details are listed in Table 1.
- Most Sold Season - First we divided the months into Spring, Summer, Fall & Winter and then for each SKU identified what session was the SKU most sold in using the SKUINFO table

- Discounted - Tagged whether a particular SKU was sold on a discount or not. The discount was identified as comparing the original price and amount and if the amount was less than the original price, we tag the feature as discounted.

The characteristics of these 5 features are summarized in Table 2.

2.2 K-prototype clustering

As we navigate through the extensive dataset from Dillard's, characterized by diverse product attributes, the k-Prototype method proves to be an indispensable tool for efficient clustering and subsequent analysis. This method combines the strengths of k-Means clustering and the k-Modes algorithm, allowing us to seamlessly handle both numerical and categorical data.

K-Prototype initializes cluster centroids with numerical values and modes for categorical features (Huang 1997). The assignment step employs a dissimilarity measure that considers both numerical and categorical distances, ensuring a holistic evaluation of data point proximity to cluster centroids. The update step recalculates centroids by computing means for numerical features and modes for categorical ones. Through iterative refinement, the algorithm converges to stable cluster assignments, accommodating the distinct characteristics of each data type.

In the context of our project, the k-Prototype method plays a crucial role in the initial phase of our model. We apply this method to five carefully selected features of each product, such as color and department, to group similar items together. This clustering step serves as a strategic precursor to the subsequent application of Market Basket Analysis (MBA), contributing to computational efficiency and the precision of our recommendations.

2.3 Market Basket Analysis

Market basket analysis is used to study retail data about items that are frequently bought together from the transaction data given, which is great for understanding customer preferences and purchasing behaviors (Hermina, 2022). While we are trying to make recommendations for items consumers may purchase based on the items they plan to purchase, market basket analysis really stands out to capture the hidden relationship inside the transaction data.

After performing K-Prototype clustering analysis with 5 chosen features, we clustered SKU into 3 different clusters that reduced the computational complexity of exploring the relationship between SKUs in the entire transaction data. By applying market basket analysis for each cluster, we could find similar SKUs that consumers are more likely to purchase together, therefore, this could give us business insights to make more revenues by offering discounts only to one item in the rules generated, giving recommendations for customers, or placing those related items close to each other in physical stores.

3. Results

3.1 Clustering results

Utilizing a clustering approach based on key features such as item discount status, seasonal popularity, return rate, size, and department, we categorized the dataset into three distinct clusters. Cluster 2 comprises 291,111 items, Cluster 1 includes 255,980 items, and Cluster 0 encompasses 29,402 items. Profile plots (Figure 2 to Figure 6) were generated to visually elucidate the characteristics and trends within each cluster.

Cluster 2 stands out with a notable percentage of discounted items and a prevalence of items most frequently bought in the summer season. Additionally, this cluster exhibits the lowest return rate among the three clusters, suggesting a higher customer satisfaction level. In contrast, Cluster 0 demonstrates the highest return rate, indicating potential areas for improvement in customer satisfaction. Cluster 1 showcases relatively larger items, although the difference is not statistically significant. Moreover, this cluster houses a higher concentration of items from department 1, known for fashion brands.

The presented findings offer valuable insights into the distinct profiles of each cluster, laying the groundwork for informed decision-making for market basket analysis.

3.2 Market Basket Analysis results

After joining the clustering results with transaction data, we grouped the transaction data into lists of SKUs for each transaction. After applying market basket analysis for each cluster, we generated 565 rules in total with support of 8%, it only looked into different combined pairs that showed up at least 8% in all the transaction dataset, we ignored the pairs that have occurrence less than 8% in the dataset since there are not enough data to conclude the relationship between SKUs in those pairs. Part of the generated results are demonstrated in Table 3.

We set the confidence to be 50% for the rules generated, for rules A->B, we only generate rules if there are more than 50% of transactions that have SKU B among the data that has SKU A. So we are quite confident about the rules we generated since we can tell there is enough data showing the relationships between the pairs we are investigating.

We can also tell from the lift values for our rules, lift measures how much more often the antecedent and consequent occur together than expected if they were statistically independent, most of our lift values are significantly bigger than 1, which also confirmed the reliability of the rules generated, therefore, it provided us with confidence for our business insights and recommendations.

4. ROI

The implementation of a Market Basket Analysis (MBA) recommendation system in our online retail operation represents a strategic foray into enhancing customer engagement. Leveraging the sophisticated capabilities of predictive analytics, this system is meticulously designed to deliver tailored product recommendations, with the dual aim of amplifying transaction values and cementing customer loyalty. The foundation of this initiative rests on a set of carefully crafted assumptions. Central to these is the projection of a nuanced yet significant elevation in the purchase rates due to recommendations – a shift from 3% to 3.1%. This seemingly modest increase is poised to exert a substantial cumulative impact on overall sales, manifesting particularly across a broad array of transactions.

Our approach to financial forecasting is anchored in the industry's consistent growth trajectory, as denoted by a stable Compound Annual Growth Rate (CAGR). This is coupled with prudent adjustments for inflation, ensuring that our ROI calculations are not only accurate but also contextually relevant, reflecting the true value of monetary figures over time. The anticipated revenue boost is directly ascribed to the advent of the recommendation system, thereby sharply accentuating its financial impact.

The investment in the recommendation system is comprehensive and multifaceted. It encompasses the initial development costs, the recruitment of top-tier data specialists for ongoing refinement of the algorithm, and significant investments in a resilient technological backbone. Further, the model accounts for ongoing operational costs, including system maintenance, iterative updates, and targeted marketing initiatives to drive user adoption. The computation of ROI, which juxtaposes the incremental revenue against the overarching investment, reveals a positive outcome. This affirmative ROI not only corroborates the efficacy of the recommendation system but also underscores the judiciousness of the investment. ROI can be seen in Figure 8.

5. Conclusion

The integration of k-Prototype clustering and Market Basket Analysis (MBA) into Dillard's online retail strategy has successfully enhanced the customer shopping experience by offering optimized product recommendations. This approach, which grouped products based on essential features and then applied MBA for detailed product association rules, increased both computational efficiency and recommendation accuracy. The methodology effectively navigated the complexities of consumer purchasing patterns, leading to improved inventory management and targeted marketing strategies.

Financially, the project yielded a positive Return on Investment (ROI), demonstrating its economic viability. The slight but significant increase in the purchase rate, from 3% to 3.1%, indicated a measurable impact on sales, supported by a stable industry growth rate and careful inflation adjustments. The balance between the initial investment in technology, expertise, and infrastructure, and the revenue generated from the improved recommendation system, underscores the strategic value of the project. This successful implementation not only boosts Dillard's current sales but also sets a precedent for future data-driven enhancements in the retail sector.

Appendix I - References

Huang, Zhexue. “A fast clustering algorithm to cluster very large categorical data sets in data mining.” *Dmkd* 3, no. 8 (1997): 34-39.

Hermina, Cecil. “Market Basket Analysis for a Supermarket.” International Journal of Management, Technology And Engineering, Volume XII (2022): 106-113.

Appendix II - Supporting Data

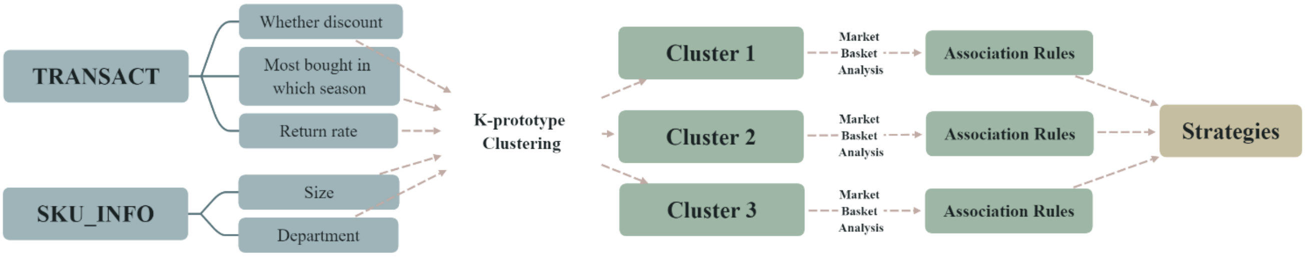


Figure 1. Central Framework

Database Diagram

PK – Primary key
FK – Foreign key

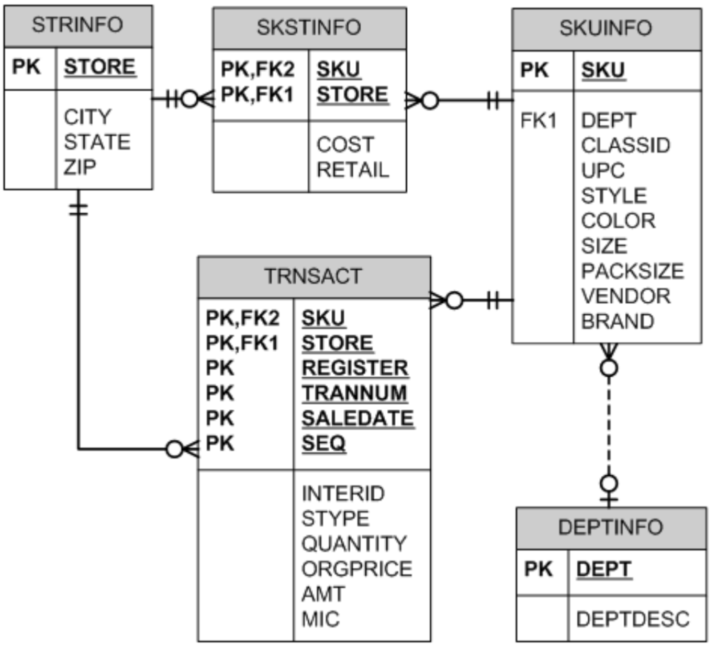


Figure 2. ERD for the Data

Class	Departments included
Class 1: Fashion Brands	R LAUREN, LACOSTE, ESPRIT, COLEHAAN, GOTTEX, DKNY, LIZ CLA, CARTIER, MOSCHINO, ST JOHN, CARTERS, C KLEIN, NO FEAR, REEBOK, ANNASUI, SPERRY,, CATALIN, CLINIQUE
Class 2: Uncertain	LESLIE, GARY F, JACQUES, CABERN, BE2, CAB, MAI, CELEBRT, BEP, 4711, CAMBIO, AUSTIN, CITIZENS, BORA, 1928, COLOR, P&Y, H SIERR, KAREN K, BLUE, FREDERI,, ELLEN T, MATTY M, CAROLE, BCH, COFFRET, LOUISVL, SIGRID O, ETERNIT, COP KEY
Class 3: Personal Name Brands	R TAYLOR, CHORUS, CAB, ONEIDA, ENVIRON, POLOMEN, GOTTEX (Possibly a duplication), COLOR (Possibly a duplication), COLEHAAN (Possibly a duplication), INVEST, NOB, BRIOSO, ECHO, LISLI

Table 1. Classification of departments

Attribute	Description	Value types	Sample data
SKU	Stock Keeping Unit number of the stock item	Integer	4757355, 2128748, ...
return_rate	Ratio of return of the item from transaction data	double	0.259259, 0.000000,…
size	The size of the stock item	text	S, L, OTHER…
dep_1	Whether belongs to Fashion and Apparel Departments	integer	0, 1
dep_2	Whether belongs to Generic or Unclear Departments	integer	0, 1
dep_3	Whether belongs to Personal or Individual Departments	integer	0, 1
most_sold_season	In which season the item is sold most from transaction data	text	Fall, Spring, …
discounted	Whether the item has discount (1) or raise price (-1) before from transaction data	integer	-1, 0, 1

Table 2. Summary of five selected features

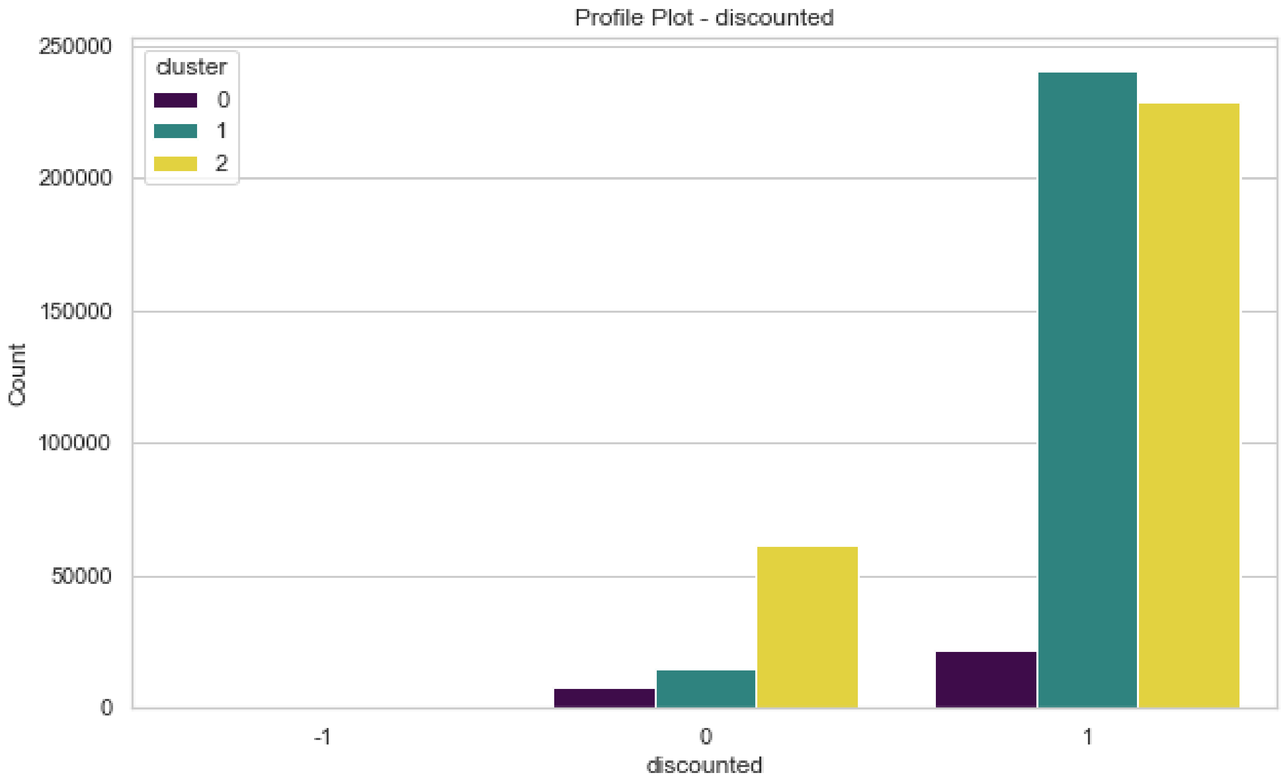


Figure 3. Profile Plot of Whether Discounted

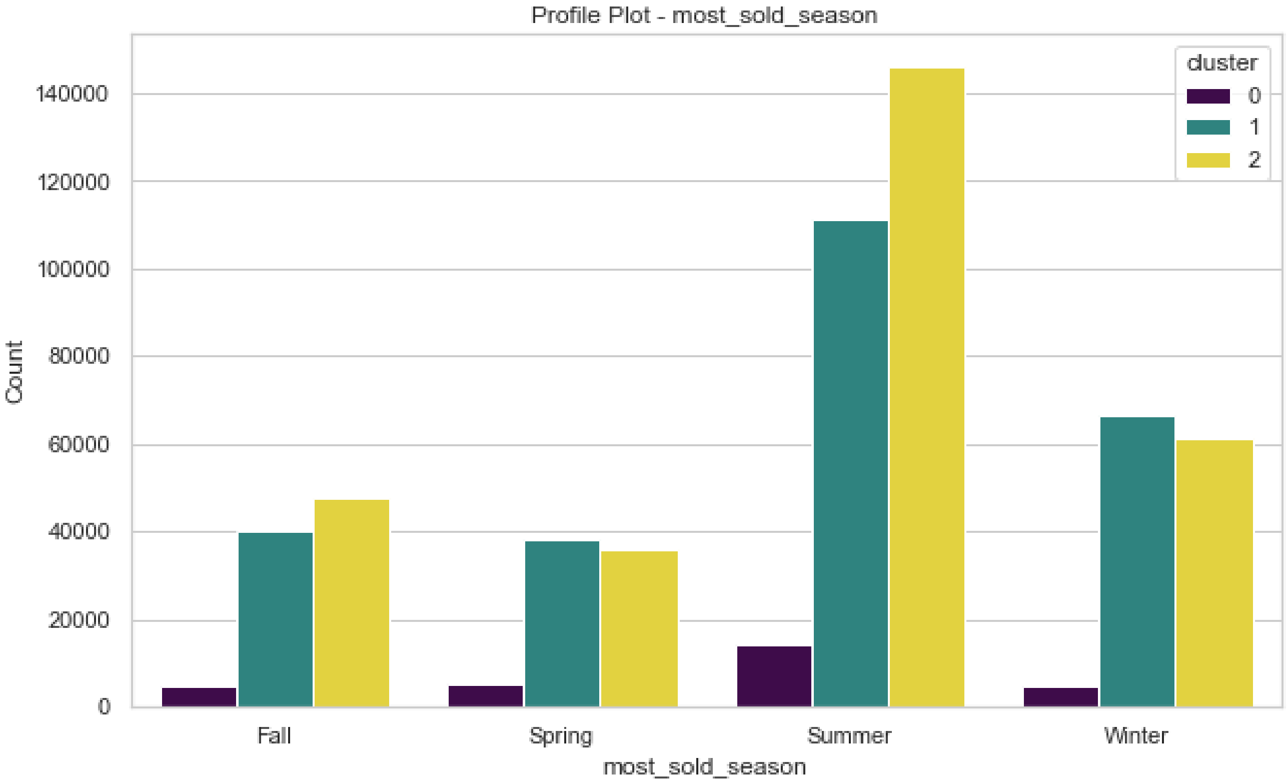


Figure 4. Profile Plot of Most sold in Which Season

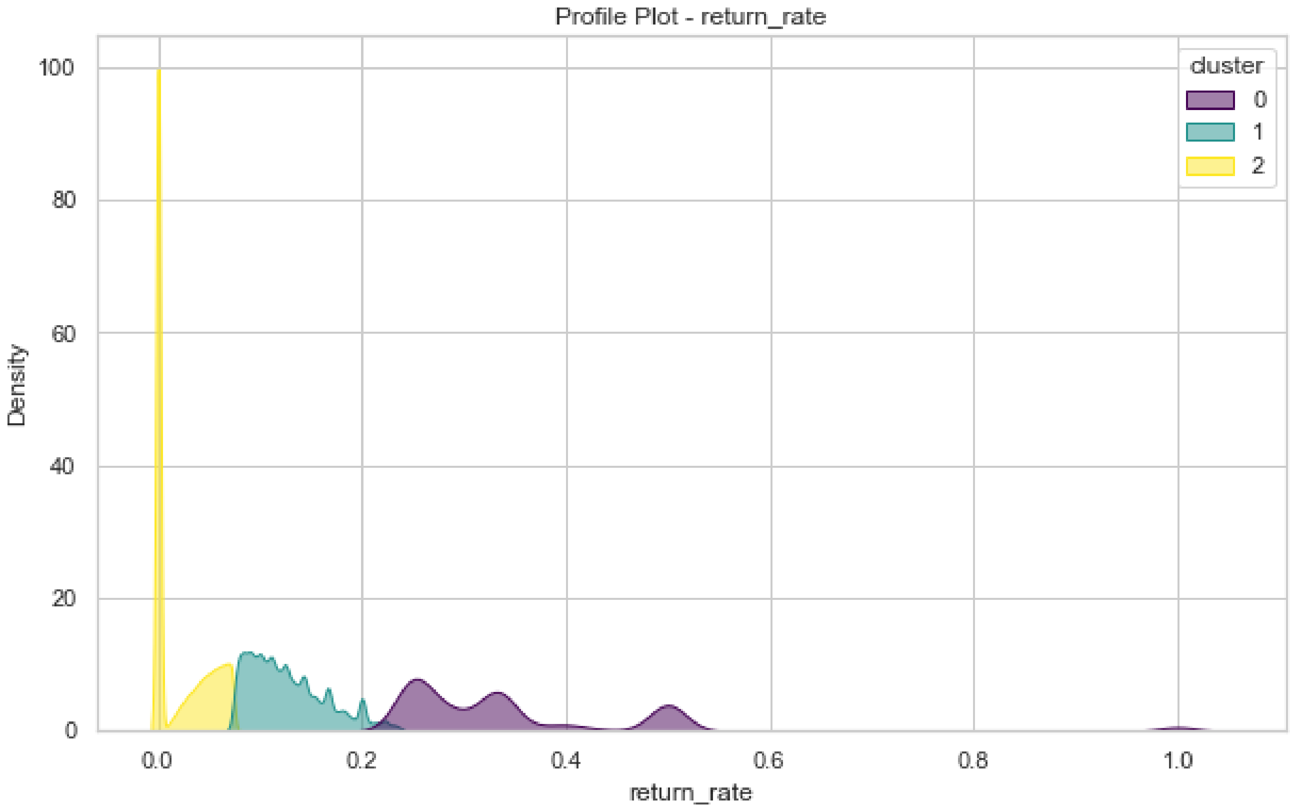


Figure 5. Profile Plot of Return Rate

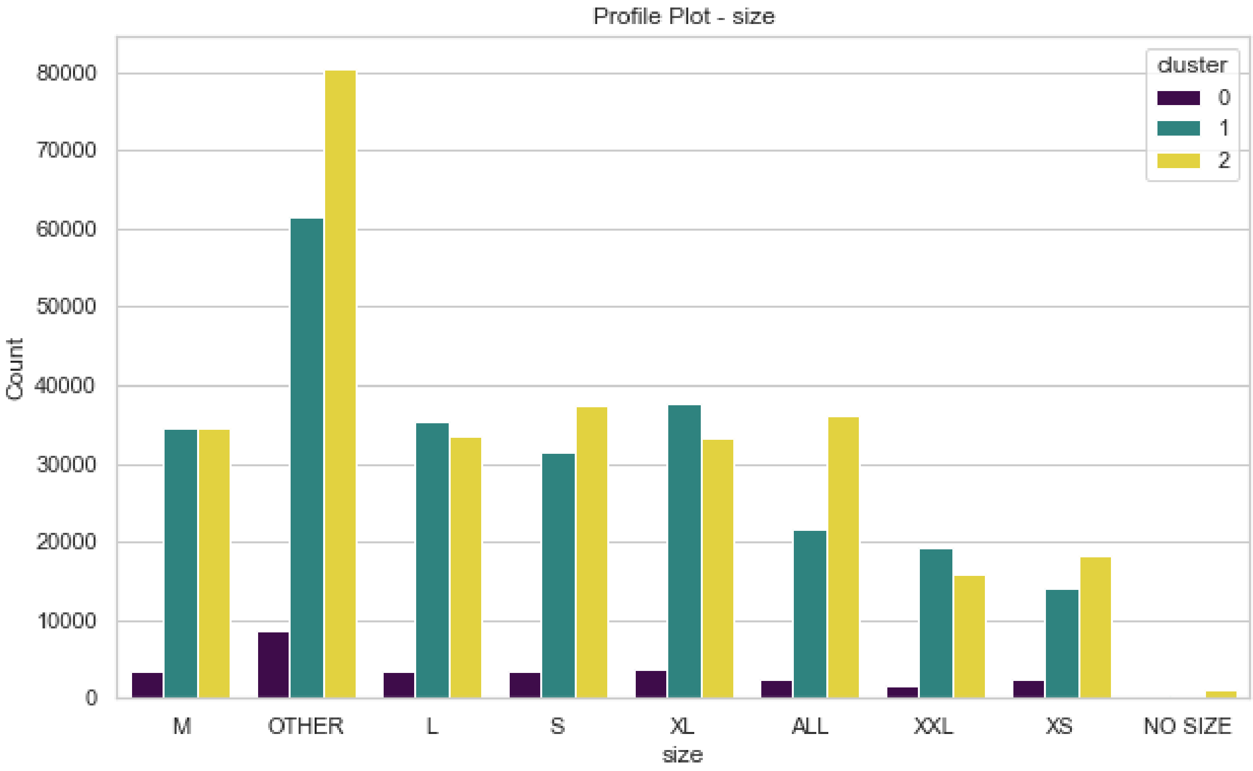


Figure 6. Profile Plot of Size

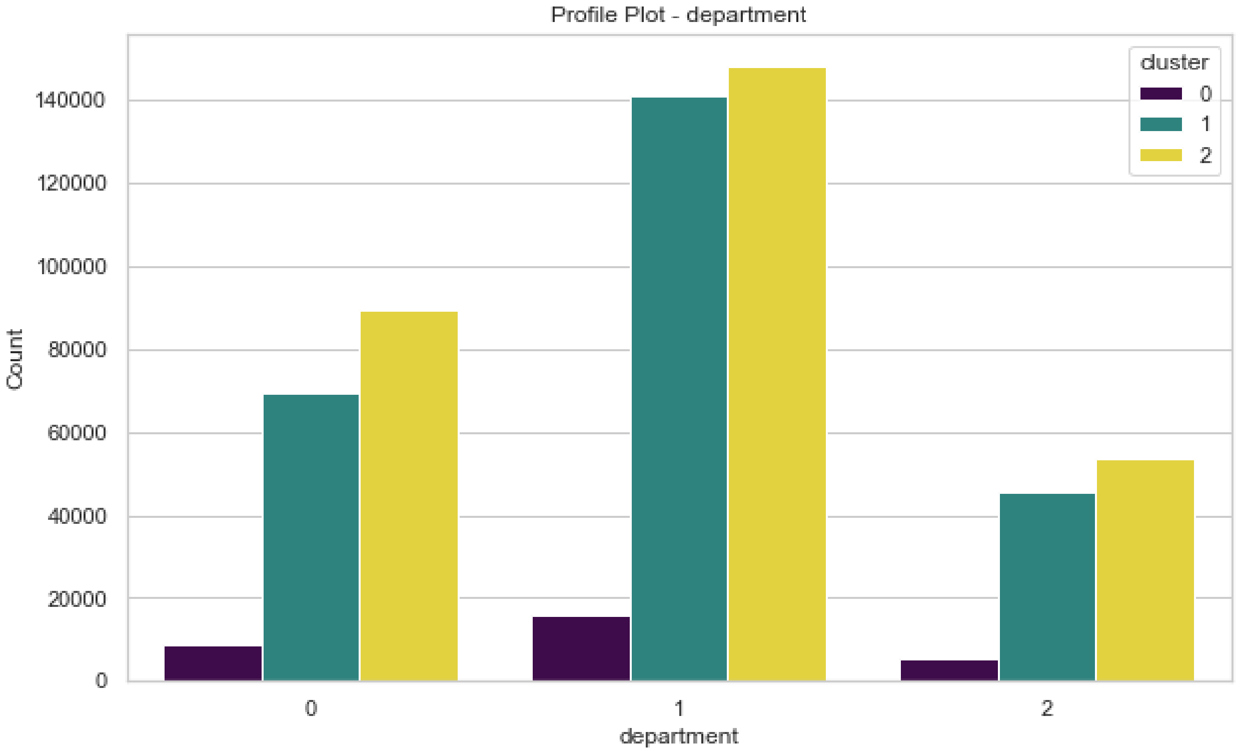
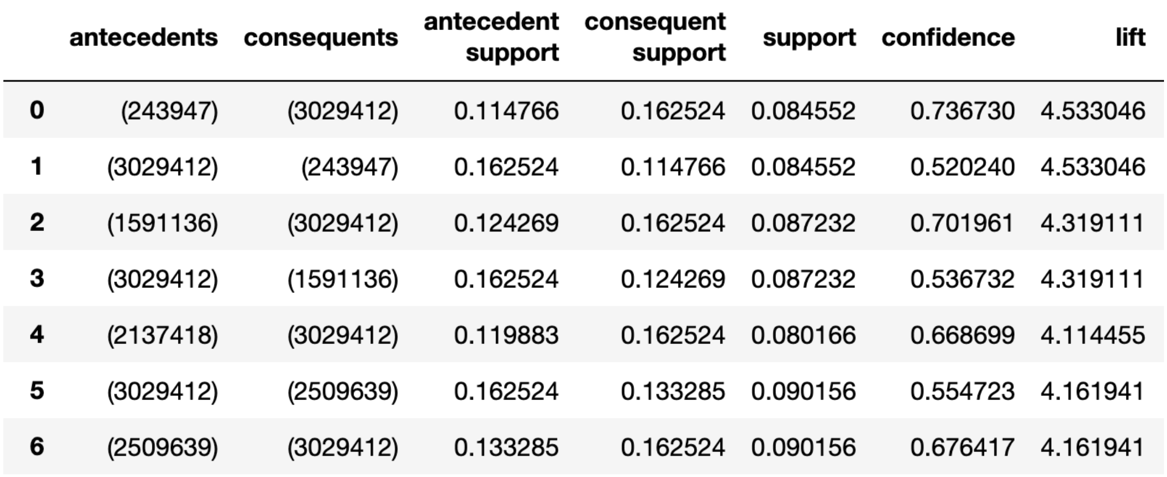
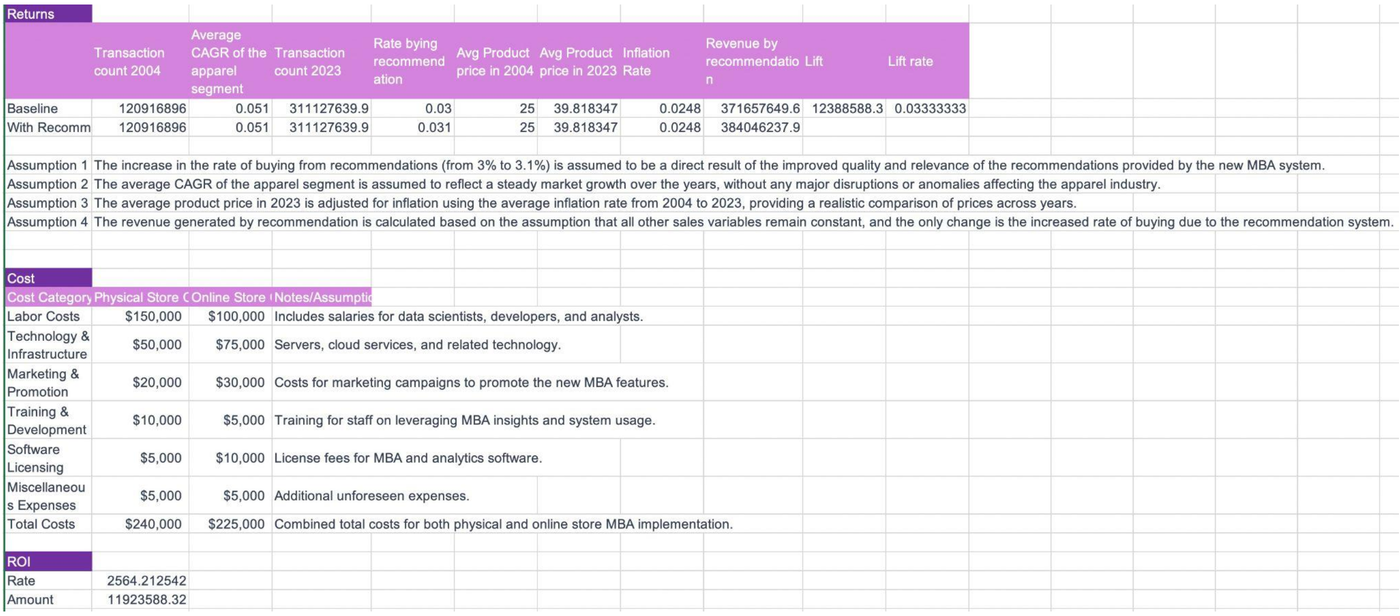


Figure 7. Profile Plot of Department



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(243947)	(3029412)	0.114766	0.162524	0.084552	0.736730	4.533046
1	(3029412)	(243947)	0.162524	0.114766	0.084552	0.520240	4.533046
2	(1591136)	(3029412)	0.124269	0.162524	0.087232	0.701961	4.319111
3	(3029412)	(1591136)	0.162524	0.124269	0.087232	0.536732	4.319111
4	(2137418)	(3029412)	0.119883	0.162524	0.080166	0.668699	4.114455
5	(3029412)	(2509639)	0.162524	0.133285	0.090156	0.554723	4.161941
6	(2509639)	(3029412)	0.133285	0.162524	0.090156	0.676417	4.161941

Table 3. Partial Demonstration of Market Basket Analysis Results



Returns

	Transaction count 2004	Average CAGR of the apparel segment	Transaction count 2023	Rate bying recommendation	Avg Product price in 2004	Avg Product price in 2023	Inflation Rate	Revenue by recommendation	Lift	Lift rate
Baseline	120916896	0.051	311127639.9	0.03	25	39.818347	0.0248	371657649.6	12388588.3	0.03333333
With Recomm	120916896	0.051	311127639.9	0.031	25	39.818347	0.0248	384046237.9		

Assumption 1	The increase in the rate of buying from recommendations (from 3% to 3.1%) is assumed to be a direct result of the improved quality and relevance of the recommendations provided by the new MBA system.
Assumption 2	The average CAGR of the apparel segment is assumed to reflect a steady market growth over the years, without any major disruptions or anomalies affecting the apparel industry.
Assumption 3	The average product price in 2023 is adjusted for inflation using the average inflation rate from 2004 to 2023, providing a realistic comparison of prices across years.
Assumption 4	The revenue generated by recommendation is calculated based on the assumption that all other sales variables remain constant, and the only change is the increased rate of buying due to the recommendation system.

Cost

Cost Category	Physical Store (	Online Store (	Notes/Assumptic
Labor Costs	$150,000	$100,000	Includes salaries for data scientists, developers, and analysts.
Technology & Infrastructure	$50,000	$75,000	Servers, cloud services, and related technology.
Marketing & Promotion	$20,000	$30,000	Costs for marketing campaigns to promote the new MBA features.
Training & Development	$10,000	$5,000	Training for staff on leveraging MBA insights and system usage.
Software Licensing	$5,000	$10,000	License fees for MBA and analytics software.
Miscellaneous Expenses	$5,000	$5,000	Additional unforeseen expenses.
Total Costs	$240,000	$225,000	Combined total costs for both physical and online store MBA implementation.

ROI

Rate	2564.212542
Amount	11923588.32

Figure 8. ROI