Detailed Version of Results

Vector CAG – Maximizing Lifetime Return Through Data-Driven Insights

Team: Vector Visionaries

Date: December 5, 2025

Project Context and Objectives

This project aims to investigate how Vector can use its historical customer and order data to improve retention and long-term revenue. Our preliminary data preparation and exploratory analysis revealed two core challenges: revenue is heavily concentrated in a small loyal group, and most customers purchase irregularly, offering limited visibility for sales planning. To support the goal to build a repeatable analytical framework that helps Vector determine whom to prioritize and when to engage, we continue to apply and refine two complementary modeling approaches. Clustering is used to segment customers and industries into interpretable groups that reflect meaningful behavioral differences. Forecasting then identifies seasonality and growth potential across industries, allowing outreach and planning to be timed more effectively. Together, these components form the foundation of a system designed to guide Vector's sales efforts with clearer signals and more actionable insights.

Results

1. Clustering Insights

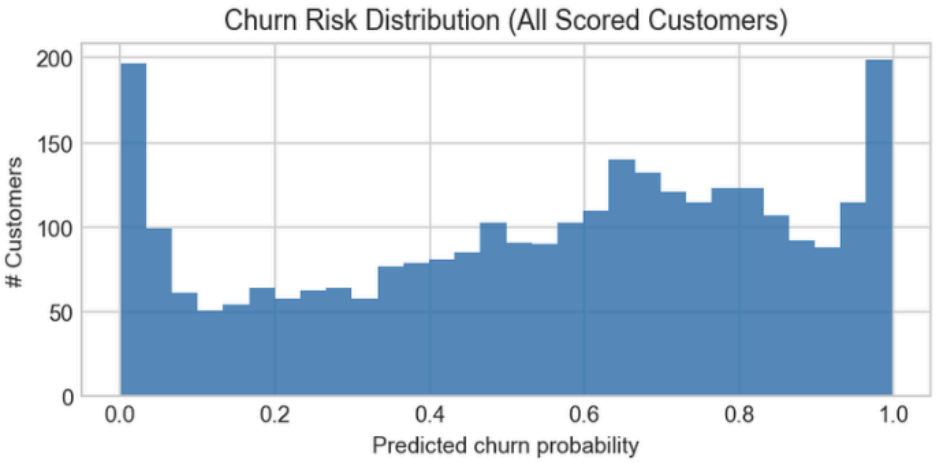
The clustering analysis reveals distinct customer segments that differ in spending behavior, platform engagement, and long-term stability. The purpose of this step is not only to classify customers into groups but also to pinpoint which clusters exhibit early signs of churn risk. These insights create a foundation to prioritize outreach and customize retention actions based on the behavioral motivations of each group.

a. Revenue Trend

One of the most critical dimensions examined across clusters is the revenue trend over time. Revenue demonstrates a generally stable and gradually increasing trajectory.

Seasonal fluctuations exist, and no structural long-term decline was observed, suggesting that retention initiatives should emphasize maintaining consistency and preventing churn rather than recovering.

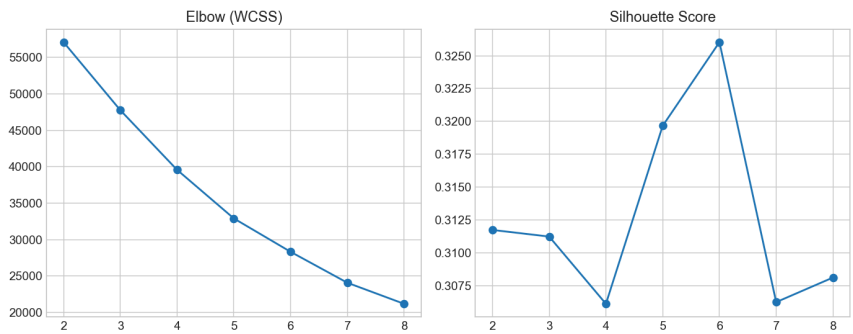
b. Churn Risk Distribution

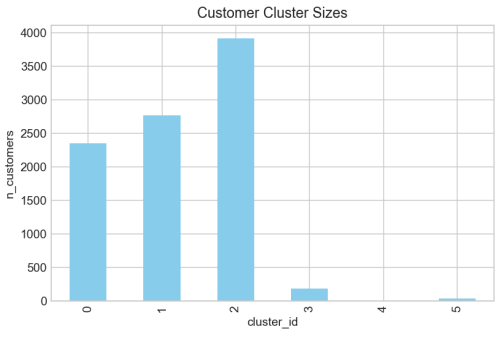


The distribution of predicted churn probabilities is U-shaped: a large portion of customers exhibit very low or very high churn likelihood, with fewer in the middle. This polarization indicates that churn is not evenly distributed across the base, suggesting segmentation-driven interventions are essential for effectively directing sales efforts.

c. Customer Segmentation

To identify meaningful behavioral segments within the customer base, we conducted an unsupervised clustering analysis using the K-Means algorithm and evaluated multiple values of k through both the elbow criterion and silhouette scores. Initial diagnostics suggested that $k = 6$ could provide the strongest cluster separation. However, when we attempted to train the model, one of the clusters consistently collapsed into an empty group, indicating that the data did not support six well-formed segments.



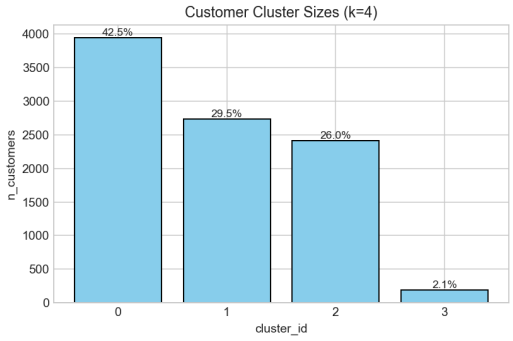


We then tested $k = 5$, which the elbow method also supported, as the WCSS curve began to flatten beyond this point, signaling diminishing returns. While five clusters were more stable overall, we still observed that one cluster was extremely small—containing only a single or near-duplicate point after scaling—making it non-actionable from a business perspective. To address this, we implemented a two-step refinement:

(1) We introduced a preprocessing check that reduces k automatically if the number of unique scaled observations is insufficient, and

(2) We ran multiple K-Means initializations with different random seeds to avoid empty clusters, falling back to a smaller k only when all attempts produced an empty group.

This procedure ensured that the final segmentation consisted of only well-populated, meaningful clusters. The remaining tiny cluster was merged into its nearest neighbor segment to preserve interpretability and operational value. Ultimately, the final set of customer clusters achieved a balance between statistical validity and business usability, forming the first layer of the segmentation logic used throughout the dashboard and playbook.



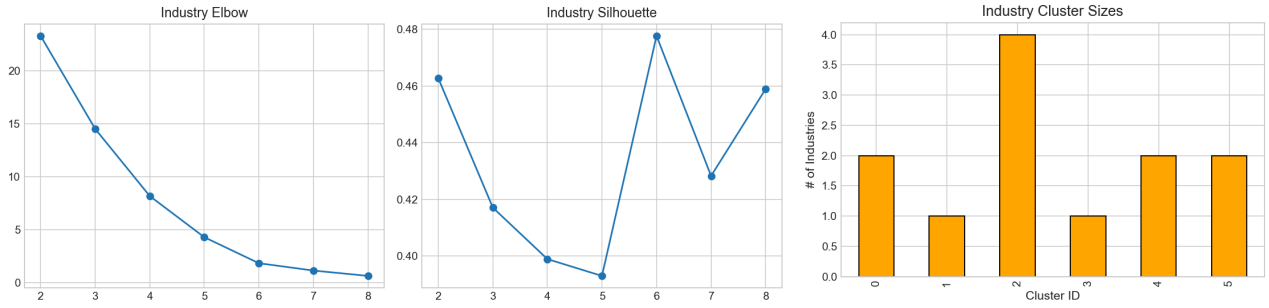
Finally, K = 4. Below is the customer cluster profile:

- Cluster 0 – Large, Low-Value & Inactive (42.5%): Low total sales, few orders, and long recency make this the largest but least engaged segment. These customers represent the long tail of inactive or low-value accounts. Reactivation campaigns or targeted promotions may help recover value.

- Cluster 1 – High-Value Loyal (29.5%): High sales, frequent purchasing, large order values, long tenure, and recent activity. This is the core, highly profitable customer base. Relationship maintenance and premium service strategies are appropriate.
- Cluster 2 – Moderate-Value Newer (26.0%): Moderate spending and frequency with shorter tenure and developing engagement. These newer customers show growth potential and may benefit from onboarding or cross-sell initiatives.
- Cluster 3 – High-Return / Unstable (2.1%): Very small group with low sales but unusually high return ratios and inconsistent activity. This cluster reflects dissatisfied or misaligned customers and warrants investigation and targeted churn mitigation.

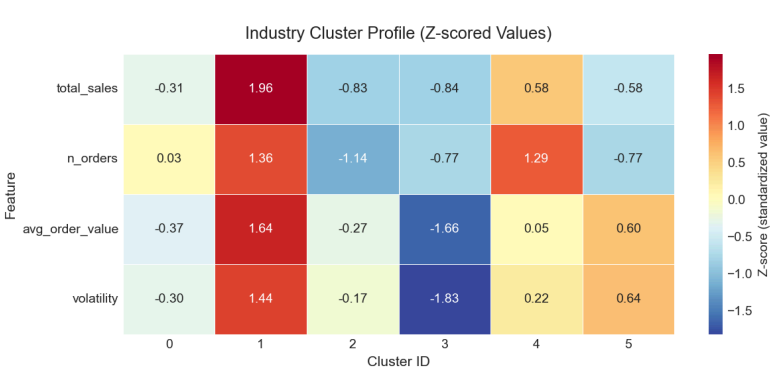
d. Industry Segmentation

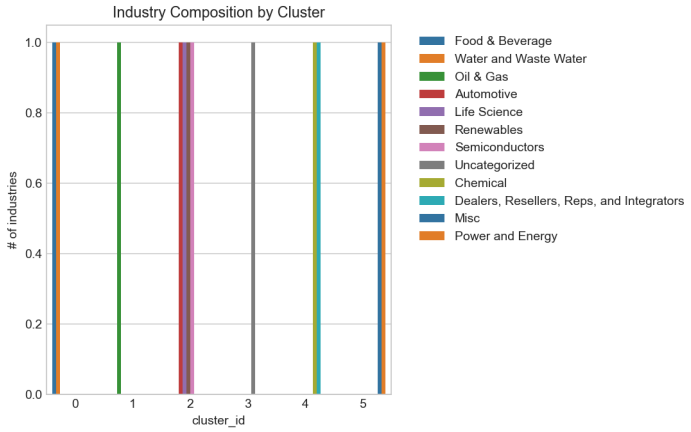
To understand structural differences across sectors, we performed industry-level clustering using K-Means applied to aggregated behavioral metrics: total sales, number of orders, average order value, and revenue volatility. As with the customer segmentation, we evaluated multiple values of k using both the elbow criterion and silhouette scores. The elbow curve shows diminishing returns beyond k = 6, while the silhouette analysis reaches its global maximum at k = 6, indicating the strongest separation among industry groups.



Unlike the customer-level data, the industry dataset consists of a small number of aggregated records, so no empty-cluster issue emerged. The model trained cleanly, and the resulting six clusters were all valid and

meaningfully populated. Cluster sizes range from one to four industries, reflecting moderate diversity across sectors.





The heatmap of standardized feature values highlights distinct patterns in scale and stability, allowing clear interpretation:

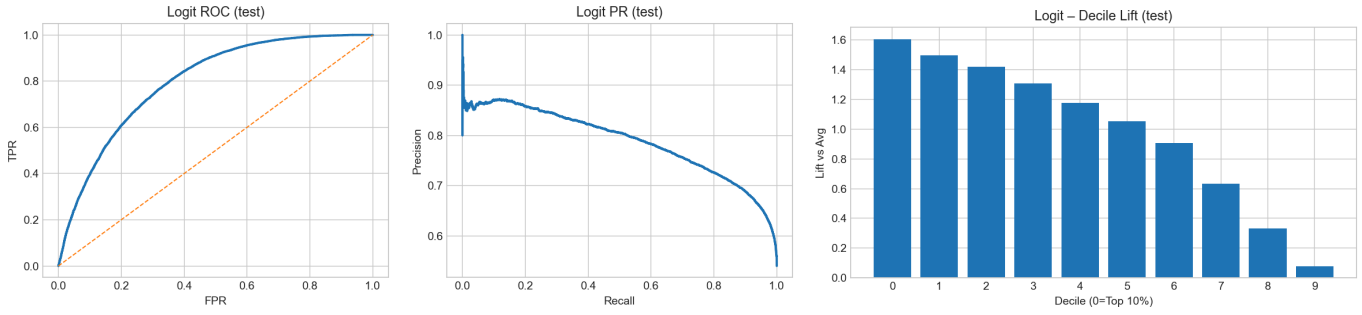
- **Cluster 1** shows extremely high total sales, order volume, average order value, and volatility — representing large-scale, project-driven sectors that generate substantial revenue but fluctuate heavily.
- **Cluster 4** has above-average sales and order counts with moderate volatility, suggesting stable high-volume industries with predictable demand.
- **Cluster 0** sits slightly below average across most features, functioning as steady mid-tier contributors without pronounced strengths or weaknesses.
- **Cluster 2** performs below average on revenue and order size, consistent with low-value, high-frequency sectors such as services or fragmented markets.
- **Cluster 3** exhibits low sales but very high volatility, indicating unstable and irregular purchasing behavior.
- **Cluster 5** shows moderate order value and moderate volatility but low overall scale, reflecting niche or emerging segments with potential upside.

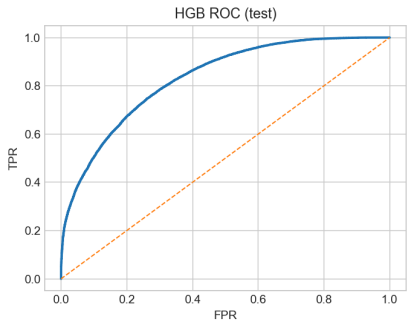
Together, these results position industries into three strategic tiers. Core strategic sectors (Clusters 1 and 4) drive revenue and should be central to capacity planning due to their volatility. Stable contributors (Cluster 0) provide baseline cash flow and forecasting stability. Emerging or fragmented sectors (Clusters 2, 3, and 5) represent either risk or growth optionality depending on strategic priorities.

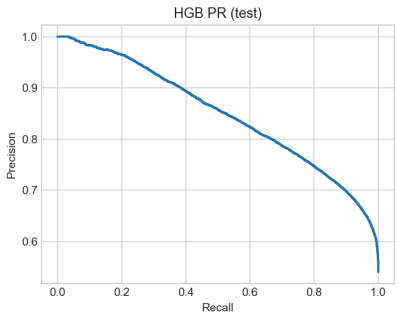
This segmentation forms the second layer of the overall analytical framework. When combined with customer clusters and individual churn risk, these industry labels help explain behavioral differences across sectors and support differentiated forecasting, scenario planning, and go-to-market strategies.

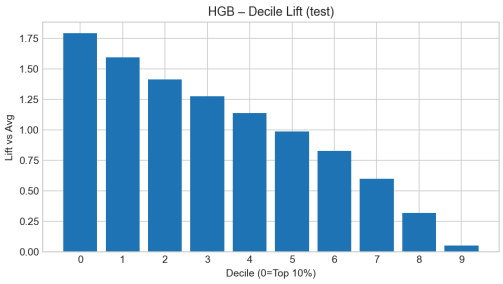
e. Model Selection for Churn Prediction

Two models were developed and evaluated for churn prediction: a Logistic Regression model (with imputation and standardization applied through a pipeline) and a Histogram Gradient Boosting (HGB) classifier. Both models were trained on the same feature set and assessed using ROC-AUC, PR-AUC, decile lift curves, and a business-oriented Top-10% targeting evaluation. Permutation importance was computed for the selected model to support interpretability.









Across all evaluation metrics, the HGB model demonstrated stronger performance, particularly in PR-AUC and high-risk ranking quality—both essential for imbalanced churn problems. HGB also produced higher decile lift in the top risk buckets, indicating more effective separation between churners and non-churners. Based on these results, HGB was selected as the final model for downstream retention analysis and risk-tier assignment.

f. Customer Cluster Playbook

- Large, Low-Value & Inactive (≈40%): High recency / low revenue; candidates for strategic reactivation.
- High-Value Loyal (≈30%): High frequency, high spend, long tenure; core strategic accounts.
- Moderate-Value Newer (≈25%): Shorter relationship history.
- High-Return / Unstable (≈2%): Volatile; requires root-cause risk mitigation.

g. Industry Cluster Playbook

- Steady high-volume sectors drive predictable revenue.
- Project-based sectors show cyclic but high-value purchasing.
- Emerging/niche sectors show irregular but promising growth patterns.

h. At-Risk Analysis

We developed an at-risk customer framework that quantifies churn exposure by combining predicted churn probability, historical sales value, and expected loss (calculated as recent sales times churn probability). We ranked full portfolio and segmented accounts into four priority tiers based on their relative risk distribution.

The tier analysis shows that churn exposure is heavily concentrated at the top of the distribution. P1 and P2 customers, representing the highest 30% of accounts, collectively contribute more than $15 million in expected loss. In contrast, P4 customers account for 40% of the base yet contribute less than 2% of the expected loss. Each risk tier was assigned a corresponding retention strategy:

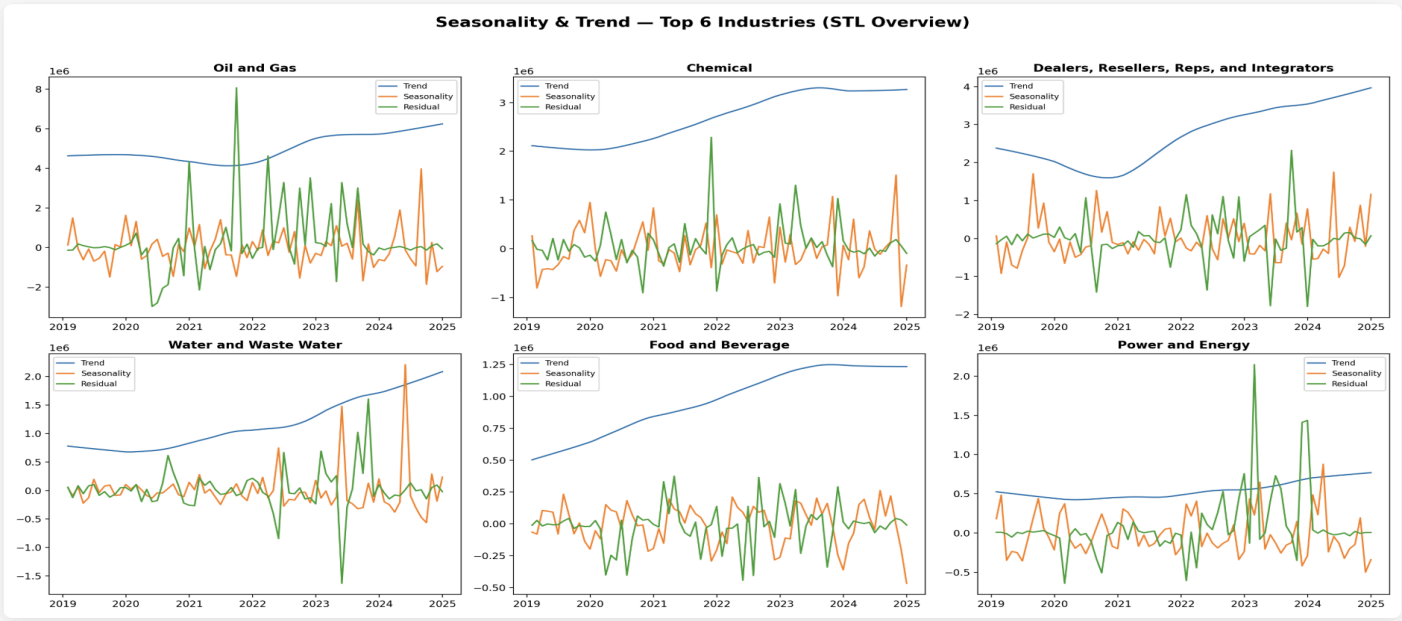
- **P1 High-touch**: Customers exhibit the highest expected loss and require immediate, high-engagement outreach to address potential issues and stabilize the account.
- **P2 Mid-touch**: These moderate–high risk accounts require structured, personalized engagement to reinforce value and reduce emerging churn signals.
- **P3 Low-touch**: The lower-risk accounts benefit from periodic, low-touch outreach that is sufficient to maintain visibility and detect early behavioral changes.
- **P4 Monitor-only**: Stable, low-risk accounts where no proactive intervention is required.

priority_band	customers	avg_churn	total_expected_loss	risk_playbook
P1 Top10%	295	97.44%	$8,986,503	High-touch
P2 10-30%	588	83.73%	$6,636,152	Low-touch
P3 30-60%	883	62.97%	$3,914,560	Mid-touch
P4 60-100%	1177	23.17%	$367,818	Monitor only

2. Industry-Level Seasonality Forecasting

The purpose of this analysis is to evaluate industry-level demand patterns and identify reliable forecasting approaches that support the company's planning needs. Based on the sectors the company considers most strategically important, we focused our work on six priority industries and applied a structured process combining decomposition, forecastability assessment, and model comparison.

a. STL Decomposition



STL decomposition reveals that most selected industries exhibit clear long-term trends, though their seasonality and residual noise differ significantly. Chemical and Dealers/Resellers show smooth upward trajectories with moderate seasonal fluctuations, suggesting consistent demand expansion. Water and Waste Water stands out with the strongest and most persistent trend, indicating accelerated growth throughout the observed period. In contrast, Oil & Gas and Power & Energy display pronounced volatility and irregular spikes, implying that their short-term behavior is dominated by external shocks. These structural differences frame each industry's inherent predictability and inform later modeling choices.

b. Forecastability Assessment

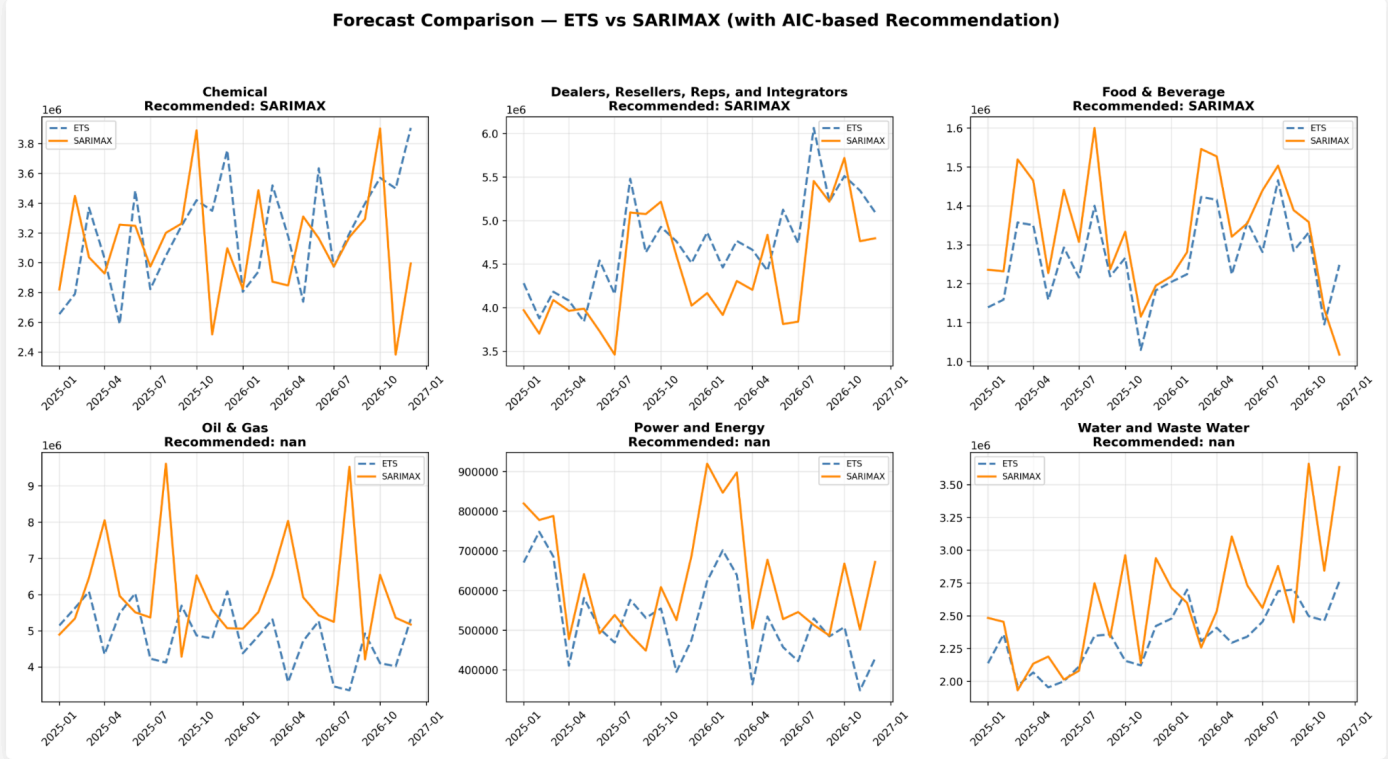
Industry Forecastability Scores

industry	n_points	seasonality_strength	trend_strength	resid_ratio	forecastable	score
Chemical	72	0.442	0.682	0.245	TRUE	0.601
Resellers, Reps, and Integrators	72	0.399	0.713	0.229	TRUE	0.599
Water and Waste	72	0.319	0.729	0.232	TRUE	0.573
Food & Beverage	72	0.291	0.739	0.243	TRUE	0.563
Misc	72	0.279	0.653	0.307	TRUE	0.511
Uncategorized	72	0.265	0.659	0.329	TRUE	0.504
Oil & Gas	72	0.282	0.257	0.578	FALSE	0.3
Automotive	72	-0.092	0.424	0.626	FALSE	0.208
Semiconductors	72	0.201	0.124	0.715	FALSE	0.187
Renewables	72	-0.025	0.29	0.714	FALSE	0.163
Power and Energy	72	0.027	0.209	0.813	FALSE	0.132
Life Science	72	0.105	0.091	0.783	FALSE	0.122

Using seasonality strength, trend strength, and residual ratio, the forecastability assessment identifies Chemical, Dealers/Resellers, Water and Waste Water, and Food & Beverage as structurally predictable, given their strong trend components and manageable noise levels. By contrast, Oil & Gas, Power & Energy, Automotive, Renewables, and Life Science show weak structural signals and elevated residual variance, yielding lower composite scores. While this ranking helps quantify predictability across all industries, the modeling scope focuses on the six company-designated priority sectors, rather than strictly on the highest-scoring ones.

c. ETS and SARIMAX Comparison

industry	model	mean_forecast	std_forecast	max_forecast	min_forecast	best_aic	volatility_ratio
Chemical	ETS	3204458	365056.5	3905038	2587851	1924.842715	0.113921
Chemical	SARIMAX	3120744	354993.3	3901263	2382754	-2.045585	0.113753
Dealers, Resellers, Reps, and Integrators	ETS	4731781	549337.4	6063341	3841329	1948.744958	0.116095
Dealers, Resellers, Reps, and Integrators	SARIMAX	4414467	633067.9	5718095	3461779	19.180031	0.143408
Food & Beverage	ETS	1263759	109817.3	1465758	1030102	1771.336382	0.086897
Food & Beverage	SARIMAX	1333445	151035	1600318	1017895	-2.508048	0.113267
Oil & Gas	ETS	4824519	815343.5	6087430	3354069	2106.580301	0.169
Oil & Gas	SARIMAX	6050272	1437837	9615724	4208186	53.921899	0.237648
Power and Energy	ETS	526752.5	109646.9	747881.5	348586.1	1835.730726	0.208156
Power and Energy	SARIMAX	627166.3	146254.5	919396.2	448529	81.386748	0.233199
Water and Waste Water	ETS	2337076	237540	2762650	1953306	1856.736306	0.10164
Water and Waste Water	SARIMAX	2599035	456483	3660387	1931092	21.411702	0.175636

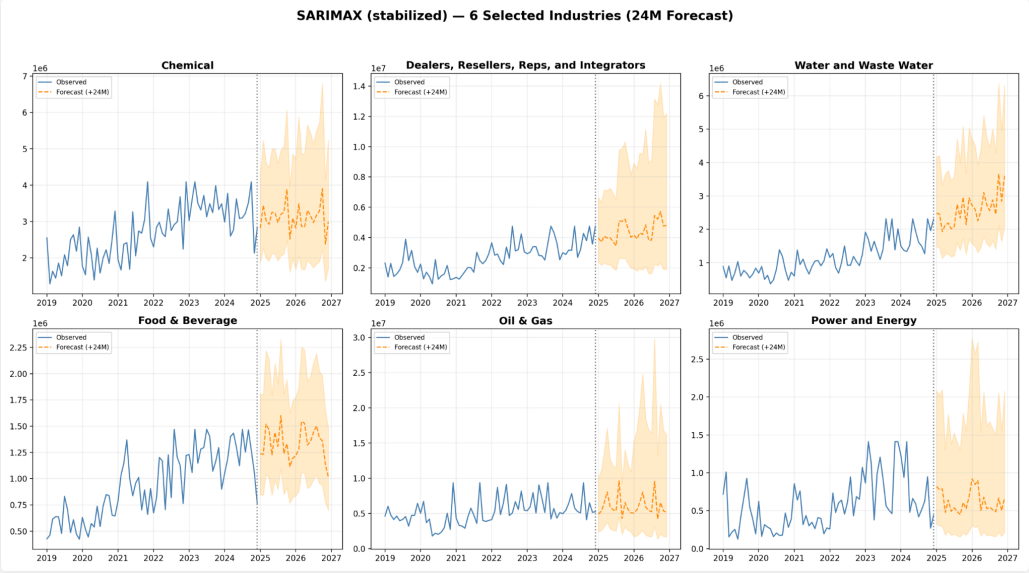


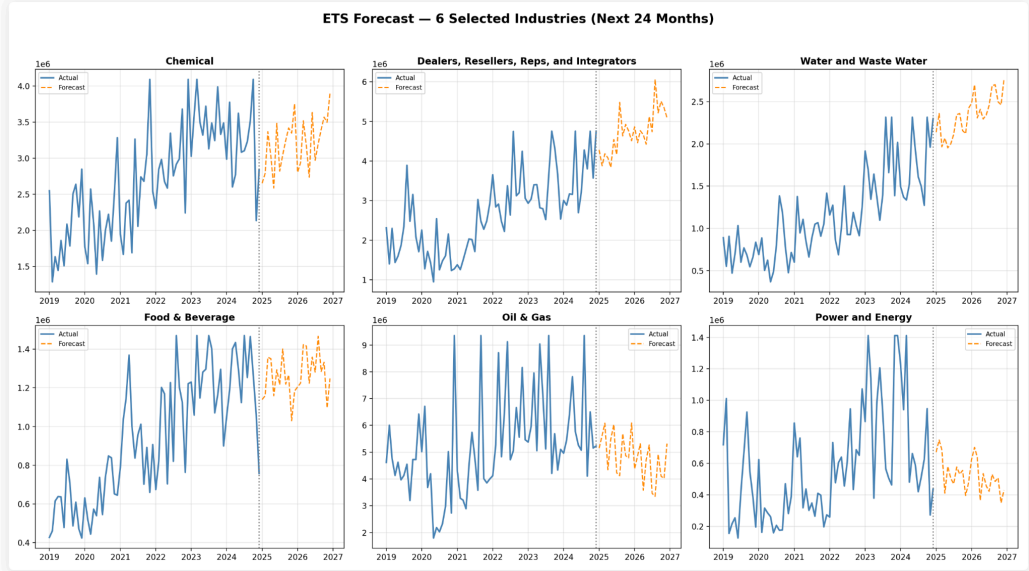
Chemical is best modeled by SARIMAX, which yields the lowest AIC and captures irregular movements effectively. Dealers and Resellers fits better with ETS due to its smoother long-term trend, while Food and

Beverage benefits from SARIMAX because it handles short-term spikes and seasonal variation well. Water and Waste Water is well represented by ETS, reflecting its strong and stable growth pattern. Oil and Gas and Power and Energy remain highly volatile but are included for strategic reasons; for both, ETS offers the most interpretable forecasts by reducing noise amplification seen in SARIMAX.

d. Final Forecast Output

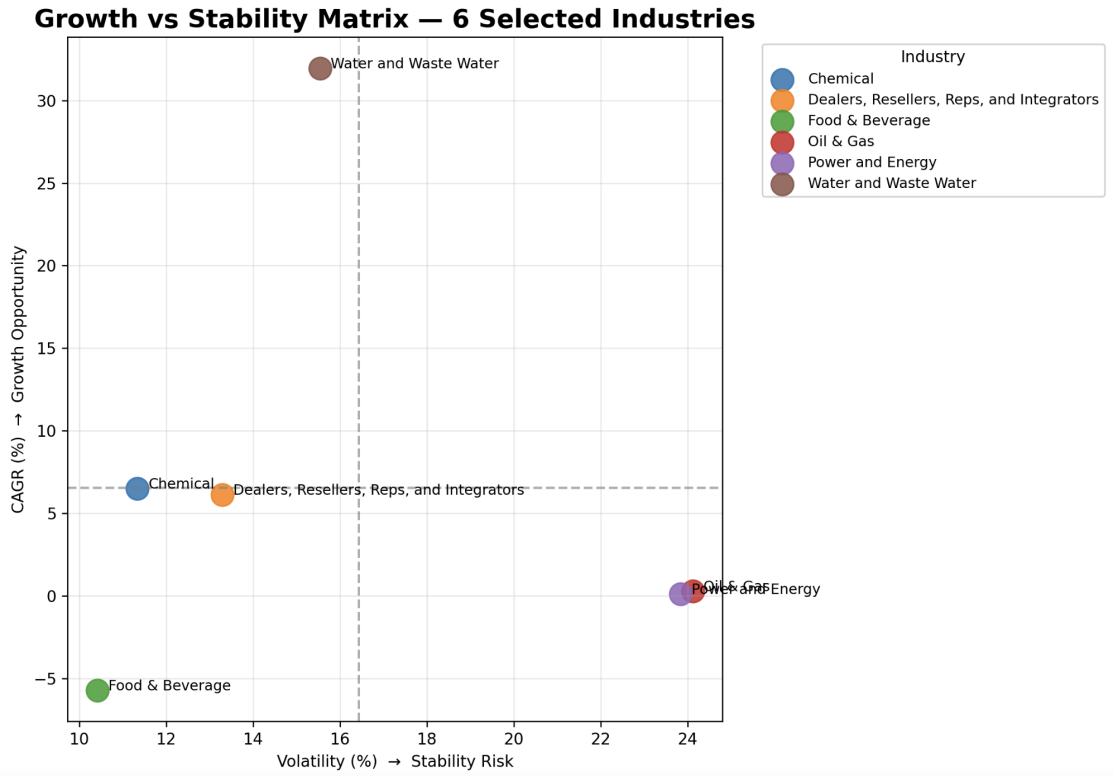
The SARIMAX-based forecasts project steady mid-term growth for Chemical and Dealers/Resellers, both maintaining upward trajectories consistent with their historical patterns. Food & Beverage shows stable seasonal cycles but slower momentum, suggesting a maturing segment. Water and Waste Water continues its strong upward climb, emerging as the most promising growth engine among the selected industries. Meanwhile, Oil & Gas and Power & Energy forecasts remain noisy and subject to large fluctuations, reflecting their structural instability. These projections offer directional





guidance while acknowledging the varying levels of model reliability across industries.

e. Growth Opportunity vs. Stability Risk



industry	CAGR (%)	Volatility (%)
Chemical	6.50	11.34
Dealers, Resellers, Reps, and Integrators	6.11	13.29
Food & Beverage	-5.72	10.42
Oil & Gas	0.27	24.12
Power and Energy	0.12	23.84
Water and Waste Water	31.96	15.54

The matrix highlights Water and Waste Water as the most attractive sector, combining the highest projected CAGR with comparatively low volatility. Chemical and Dealers/Resellers occupy moderate-growth, moderate-risk positions, offering steady returns and predictable demand. Food & Beverage shows stable but weak near-term growth, reflecting limited upside potential. Oil & Gas and Power & Energy fall into the low-growth, high-volatility quadrant, signaling heightened operational and forecasting risk. This framework

enables more informed allocation of resources, balancing growth potential against the uncertainty inherent in each industry.

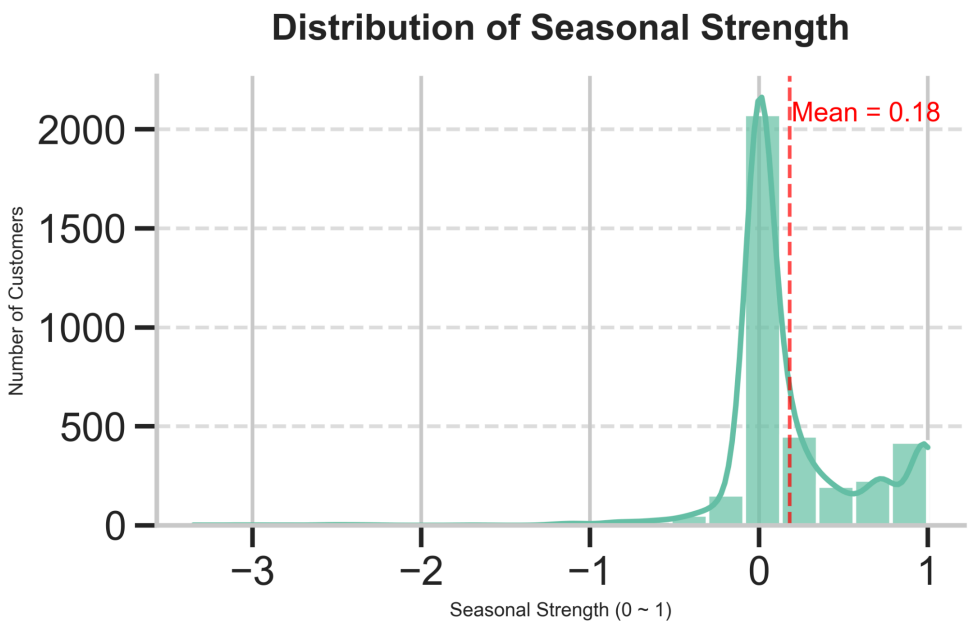
The table further reinforces the matrix patterns by showing the exact growth and stability values for each industry. Water and Waste Water stands out with a CAGR of 31.96% alongside moderate volatility of 15.54%, confirming its position as the strongest long-term opportunity. Chemical and Dealers / Resellers deliver steady mid-single-digit growth with volatility in the low-teens, matching their roles as reliable, moderate-risk sectors. Food and Beverage shows the lowest volatility at 10.42% but a negative CAGR of −5.72%, signaling stable yet declining demand. Oil and Gas and Power & Energy exhibit minimal growth—0.27% and 0.12% respectively—paired with the highest volatility in the dataset, placing them firmly in the low-growth, high-risk quadrant and requiring caution in planning.

3. Customer-Level Growth Potential Forecasting

This section presents the analytical results from the customer-level forecasting component of the project. The overall results combine three elements: customer seasonality assessment, forecasting model validation, and the subsequent segmentation and scenario evaluation. The modeling stage establishes the empirical foundation for these downstream analyses by identifying structural patterns in purchasing behavior and selecting appropriate forecasting methods for each customer.

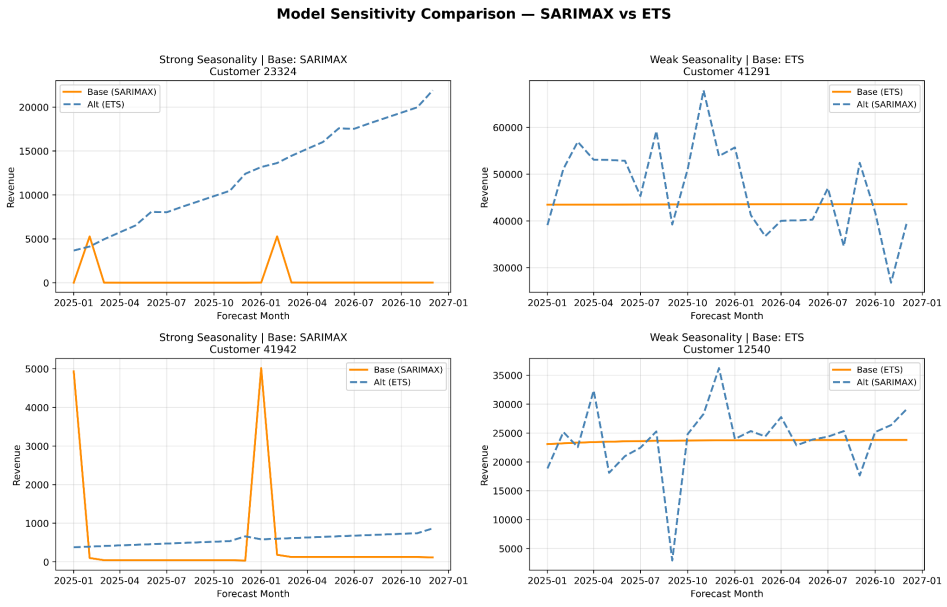
a. Seasonality Assessment

We first applied STL decomposition to measure trend, seasonality, and residual noise for each customer. The results indicate that the majority of customers exhibit weak or irregular seasonality. The average seasonal strength is approximately 0.18, which suggests that most customers do not follow predictable purchasing cycles. A smaller subset displays meaningful recurring patterns that likely reflect maintenance schedules or fiscal-year timing.



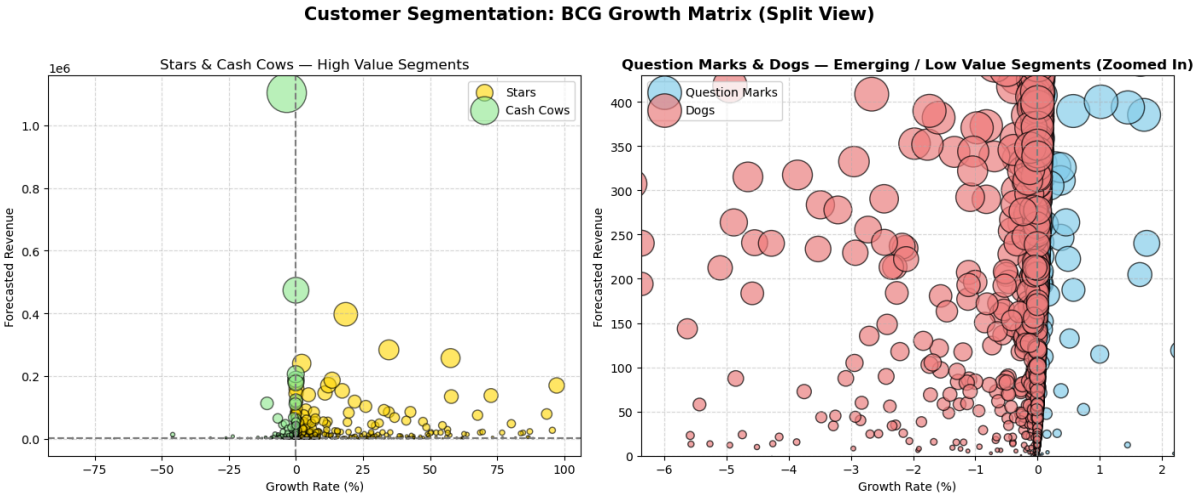
b. Model Assignment and Performance

Based on the observed seasonality patterns, two forecasting approaches were evaluated: ETS for stable or trend-driven customers and SARIMAX for customers with strong seasonal behavior. We compared model performance using backtests, focusing on forecast stability and the ability to represent recurring patterns. The majority of customers were best modeled by ETS, which generated smooth and interpretable forecasts for low-variance series. SARIMAX performed better for customers with strong seasonality by capturing periodic peaks that ETS tended to underestimate.



To validate the differentiation logic, we performed side-by-side forecast comparisons for selected customers. Among strong-seasonality customers, SARIMAX preserved the cyclical amplitude while ETS produced unrealistic linear trends. For weak-seasonality customers, ETS generated stable projections, while SARIMAX overreacted to noise and produced volatile forecasts. The validation confirms that tailoring the model to each customer's behavioral structure improves forecast reliability and interpretability.

c. BCG Matrix



To translate individual customer forecasts into actionable strategic direction, we applied the BCG Growth Matrix using two core indicators: forecasted or historical revenue growth rates and the latest annual revenue contribution. This segmentation reveals that the company's customer base is structurally diverse, with clear clusters of high-value, high-growth customers contrasted against accounts with weaker or irregular patterns.

The **Stars** segment consists of high-growth, high-value customers who consistently expand their spend and demonstrate reliable commercial behavior. These customers form the backbone of forward growth, and their purchasing patterns indicate strong alignment with the company's offering. They are well-suited for deeper commercial engagement, co-planned initiatives, and prioritized resource allocation.

Cash Cows, in contrast, generate substantial and stable revenue but exhibit limited growth momentum. These accounts typically reflect mature relationships or established industrial usage. Their value lies in stability and predictability; they provide the dependable revenue needed to balance volatility in other segments. For this group, the emphasis should be on maintaining operational efficiency, protecting margins, and enhancing service attach rates rather than pursuing aggressive expansion.

The **Question Marks** represent emerging customers whose purchasing volume remains modest but whose growth trajectory is positive. These accounts are highly sensitive to commercial activity and often respond strongly to targeted programs such as promotions, bundling, or tailored account management. Their future value hinges on timely and structured interventions capable of converting early growth signals into sustained demand. Under supportive conditions, these customers have the potential to evolve into future Stars.

Finally, the **Dogs** segment includes customers with low value and low or negative growth. Their spending is typically inconsistent or declining, and they contribute minimally to the overall revenue base. These accounts require minimal active management and should be maintained primarily where necessary for strategic coverage, geographic presence, or contractual obligations.

Overall, the BCG analysis highlights a clear hierarchy within the customer base. Stars and Cash Cows anchor current performance, while Question Marks determine the long-term upside. Dogs are best managed through automated or low-touch strategies. This segmentation establishes the foundation for more targeted and efficient commercial planning.

d. Scenario Simulation

To understand how customer value may evolve under different market and execution conditions, we conducted scenario simulations across three forecast paths: base, high, and low. These scenarios incorporate the model

assignments from ETS and SARIMAX and amplify or depress demand signals to mimic realistic commercial environments.

The Base Case reflects the most likely progression of customer revenue based on a combination of historical patterns, trend signals, and identified seasonality. Under this scenario, Stars continue to grow steadily, Cash Cows remain stable, and Question Marks show moderate but uncertain expansion. This scenario supports annual budgeting and

In the High Case, revenue is lifted by stronger sales execution, favorable market conditions, or improved customer engagement. Here, Stars accelerate meaningfully, and Question Marks show the most dramatic gains. Their sensitivity to uplift indicates that they are highly responsive to proactive commercial interventions, making them a critical target for growth initiatives. Even modest improvements in attention, marketing, or supply coordination can produce outsized returns in this group.

The Low Case simulates downside conditions such as market slowdown, cautious purchasing behavior, or reduced execution intensity. In this environment, Stars soften slightly but remain resilient, while Question Marks experience noticeable declines as their emerging growth signals dissipate. Dogs drop further, and Cash Cows maintain their role as the stabilizing core of the portfolio.

Together, these scenarios paint a clear picture of risk and opportunity: Stars remain reliable under all conditions, Question Marks offer the highest upside in favorable environments, and Cash Cows protect the downside by supplying consistent contributions.