Predicting Customer Reorders: Data-Driven Sales Optimization
Environmental Products and Accessories, LLC. (EPA Sales)

DATA 5200 Advanced Data Analytics
Valentina Nguyen
Arnoldo Ruiz
Corey Thompson

Abstract

Environmental Products & Accessories, LLC (EPA Sales) is a manufacturer and reseller of vacuum truck parts. Despite its industry presence, EPA Sales currently lacks outbound sales efforts, limiting its ability to anticipate customer needs and encourage repeat purchases. This project aims to predict customer reorders and reorder timing while identifying high-value customer segments to inform sales and marketing strategies.

Using historical 2024 transaction data, we applied K-Means Clustering to refine customer segmentation based on purchasing behavior, Random Forest to determine feature importance and predict if a customer will repurchase within 30 days, association rule mining to look for interesting and useful relationships between items, and the Cox Proportional Hazards model to estimate the likelihood and timing of future purchases.

We analyzed join-level customer, transaction, and product data to uncover purchase behaviors, product performance patterns, and time-based trends. Using K-means clustering, we identified three customer segments with distinct spend and frequency profiles; a random forest classifier (AUC > 95%) effectively predicted 30-day reorders; association rule mining revealed complementary product bundles; and a Cox proportional hazards model highlighted a critical repurchase window within 50 days. Our findings inform segmentation-driven marketing, inventory optimization, and early lifecycle outreach strategies to boost retention and revenue.

Keywords: *Customer Reorder Prediction, Sales Optimization, Survival Analysis, Cox Proportional Hazards Model, Customer Segmentation, K-Means Clustering, Purchase Behavior Analysis, Outbound Marketing Strategy*

Introduction

Environmental Products & Accessories, LLC (EPA Sales) is a leading manufacturer and reseller of vacuum truck parts, offering a diverse selection of jetting and vacuum components at competitive prices. Unlike many suppliers that focus solely on specific product lines, such as only nozzles or only Aquatech®[1] parts, we provide a broad range of components across multiple industries, making us a go-to source for various equipment needs at a faster turnaround time.

Our current market segments include:

- Industrial and Commercial Contractors: These businesses typically perform services for private enterprises or larger public projects, such as hydro-excavation, drain cleaning, and industrial waste removal.
- Municipalities and Public Works Departments: These organizations regularly use jetting and vacuum trucks to maintain sewer systems, stormwater drains, and other infrastructure.
- Environmental Services and Waste Management Companies: These companies specialize in cleaning up hazardous and non-hazardous waste, maintaining septic systems, and performing environmental remediation tasks.
- OEMs (Original Equipment Manufacturers): OEMs in this industry design and build vacuum trucks, hydro excavation trucks, jetters, and related equipment. Examples include Vactor®, Vac-Con®., and GapVax®. [2]
- Equipment Dealers: These dealers sell, lease, or service vacuum and hydro excavation trucks to end users.

These segments are categorized into customer profiles[3]:

- Level 1 (Industrial and Commercial Contractors, Municipalities and Public Works Departments): Small- to mid-sized cities (20K–250K population), and contractors operating fewer than 20 trucks.
- Level 3: Bigger cities (>260K population) that rely on equipment dealers to make purchases.
- Level 5: Industrial-level contractors with large fleets[4] (60+ trucks).
- Level 6: Independent or regional parts and equipment dealers.
- Level 7: OEMs and large-scale dealers.

---

[1] Aquatech® is a prominent U.S.-based manufacturer specializing in high-performance sewer cleaning and hydro-excavation trucks. With over 50 years of experience, Aquatech has established itself as a leader in crafting innovative and reliable equipment solutions for municipalities and specialty contractors worldwide .

[2] Vactor®, Vac-Con®, and GapVax® are U.S.-based manufacturers specializing in industrial vacuum and hydro-excavation equipment. They design and produce a range of machinery used in sewer cleaning, debris removal, and underground utility locating.

[3] In later sections of this report, Level 1 and Level 3 customers will be consolidated under the Level 1 designation for simplification.

[4] Fleets refers to the number of operational vacuum trucks, hydro-excavators, or similar equipment owned or managed by a contractor or municipality.
Small fleets: fewer than 20 trucks (typical of Level 1 customers)
Large fleets: 60 or more trucks (typical of Level 5 customers)

EPA Sales currently has the strongest hold in the Level 1 and Level 3 market segment with a 51% market share in a $10M market. Our goal is to break into the Level 5, 6, and 7 markets, as we currently hold approximately 4% of the market share in a $190M market.

Currently, the company lacks outbound sales efforts, limiting its ability to anticipate customer needs and encourage repeat purchases. By implementing proactive sales and marketing strategies, EPA Sales can strengthen customer relationships, reduce missed opportunities, and expand its market presence.

Understanding customer segmentation and predicting repurchase is an increasingly more common practice to help businesses improve their retention and outreach strategies. Prior research in marketing analytics highlights key areas that reflect our goals of reorder modeling, customer segmentation, and survival analysis in the context of B2B and e-commerce.

Reorder modeling has been widely used in e-commerce and supply chain logistics to anticipate repurchases, to keep appropriate inventory, and reduce lead times. A study demonstrates utilizing machine learning models like logistic regression and random forests to analyze historical data to classify whether a customer will repurchase within a specified window. By evaluating these models, the research identifies which algorithm offers the highest accuracy in forecasting future purchases. The findings suggest that machine learning can effectively model customer behavior, aiding businesses in making informed decisions regarding inventory management and targeted marketing strategies.[5]

Customer segmentation is a long-standing marketing technique used to determine personalized marketing strategies. K-Means Clustering has been a popular method for discovering hidden patterns in consumer behavior. Jia, Y., & Wang, J. (2019) article uses the Cox Proportional Hazards Model in eCommerce to forecast when a customer is likely to return. Survival models are particularly effective when used in conjunction with clustering to distinguish different purchasing trends across customer segments, a strategy adopted in this project to balance interpretability with predictive power. [6]

Retention rates vary widely across industries. According to Shopify (2025), the average customer retention rate for manufacturing companies is 67%, while for e-commerce businesses, it is significantly lower at 33%. EPA Sales operates under both manufacturing specialized parts and selling them on an online platform. Its customer retention challenges reflect the dynamics of both industries. This emphasizes the importance of implementing predictive analytics and targeted outbound strategies to improve retention, particularly among high-value customers. An example of this is Udaan's predictive buying approach study in 2024. It is a B2B e-commerce platform in India that implements predictive models to predict customer order patterns. This

---

[5] Emre Deniz, E., & Bülbül, S. Ç. ˇGlu. (2024). "Customer Purchase Prediction Using Supervised Machine Learning." ADBA Information Technology and Publishing Limited Company, 1(1). https://doi.org/10.69882/adba.iteb.2024071 https://www.researchgate.net/publication/382768389_Predicting_Customer_Purchase_Behavior_Using_Machine_Learning_Models

[6] Jia, Y., & Wang, J. (2019). Customer revisit prediction using Cox proportional hazard model with deep learning. *IEEE Access, 7*, 39452–39461. https://doi.org/10.1109/ACCESS.2019.2906433

resulted in a 3x increase in customer order rates, which demonstrates the effectiveness of integrating machine learning techniques to enhance customer engagement and retention.[7]

To enhance customer retention and drive revenue growth, EPA Sales aims to predict if a customer will reorder within a 30-day threshold and leverage historical data to refine customer segmentation. This predictive capability will determine if a customer will repurchase and when they will return.

A "reorder" or "repurchase" is any subsequent order placed by the same customer following a prior transaction, regardless of whether the exact product is repeated. This reflects the reality of our customer base, which includes both B2B and B2C buyers. However, our focus is on high-revenue B2B customers, identified as Priority Access segments (Levels 5, 6, and 7), who tend to place higher-value and operationally-driven orders. These customers may not always reorder the same product, but often maintain a consistent purchasing rhythm due to ongoing equipment or inventory needs.

Purchase behavior includes variables such as time between orders, order frequency, total amount spent, and number of SKUs per transaction. These variables help inform both the likelihood and timing of future reorders, with the ultimate goal of targeting retention strategies and outbound marketing efforts toward high-value, repeat buyers.

This approach enables more precise outbound marketing, improved inventory forecasting, and a streamlined customer experience tailored to high-value buyers.

Data Collection & Preprocessing

The data for this project comes directly from EPA Sales, which contains customer transaction records. The data consists of Shopify sales data that includes 2024 transaction details.

Key Observations:

- The dataset consists of transaction details, customer information, and product listings.
- The data set includes returns, unpaid transactions, and test transactions on top of regular transactions.
- The dataset consists of multiple attributes across various data types, including strings, floats, integers, and objects.
- The *trans_details* table is itemized per product, meaning that each order may appear multiple times, depending on the number of products included in the order. The *transactions* table captures the overall sale price of the order, while the *trans_details* table breaks down the total amount per product within each order.

[7] De, T. S., Singh, P., & Patel, A. (2024). A Machine learning and Empirical Bayesian Approach for Predictive Buying in B2B E-commerce. *The 8th International Conference on Machine Learning and Soft Computing (ICMLSC 2024)*, 17–24. https://doi.org/10.1145/3647750.3647754

- The *trans_details* table will be merged with the *customers* table for further analysis. We can map *trans_details* using the order number as the primary key. Selecting a unique identifier for orders was between “Order Name” and “Order ID”. “Order Name” was chosen as the unique identifier since Shopify’s search engine uses “Order Name” to retrieve orders. This ensures easy lookup and troubleshooting directly within Shopify when needed.

Several preprocessing steps were applied across the *Transactions*, *Transaction Details*, *Customers*, and *Products* tables to prepare a clean and reliable merged dataset for analysis. These steps were essential to improve data consistency, accuracy, and overall usability:

Dropping Unnecessary Columns

Irrelevant and redundant fields were removed to streamline the dataset and focus on the most informative attributes:

- *Trans_Details* Table
  - Columns such as *Kind*, *Gateway*, *Card Type*, *Payment Method*, and *Currency* were dropped. Additionally, failed and pending orders were excluded to ensure that only completed transactions were analyzed.
- *Customers* Table
  - Columns related to marketing consent (*email/SMS marketing status*), *phone numbers, customer notes, tax-exempt status*, and all address fields besides *State* and *Zip Code.*
- *Products* Table
  - Metadata, SEO attributes, inventory tracking details, product descriptions, vendor information, and international pricing fields. Additionally, the *Product Name* was excluded since it did not effectively differentiate products the way SKUs do, such as indicating specifications like GPU or PSI[8].

Filtered Out Invalid Rows

We removed rows that could distort the analysis or did not reflect real customer behavior:

- Transactions with negative values (representing returns)
- Transactions with zero total sales, which typically indicate free shipping, canceled orders, or test orders
- Transactions with zero quantity, which reflect manual adjustments to sales orders due to mistakes and errors
- Transactions from test accounts using “*bogus*” transaction types and using *@epasales.com* email domains
- Customers with no email address or 0 purchase order quantity
- Products with null *Variant SKU* were excluded to ensure all products have a valid SKU.
- Inactive products were retained, as historical transactions tied to old SKUs may still offer valuable insights

[8] GPU (Gallons Per Unit) measures water volume output, whereas PSI (Pounds per Square Inch) indicates the pressure of that water flow.

- Only 2024 transaction data was kept; records from 2023 and 2025 were excluded to maintain temporal consistency

Renamed Columns for Clarity

Renamed columns for clarity and used underscores for readability and consistency. Columns were also shortened for plotting purposes.

Standardized Data Types

- Converted *Order_Date* from object to datetime for accurate filtering and time-series operations
- Reformatted dates from YYYY-MM-DD to DD-MM-YYYY
- Changed *Order_Num* to integer type and removed any prefix symbols (e.g., #)

Customer Segmentation Adjustment

Initially, we wanted to use the *Tags* variable to identify levels 5, 6, and 7 and utilize that to aggregate by the interest level. However, tags are only assigned when a customer creates an online account, meaning they are absent for a large portion of transactions. As a result, relying on tags would provide an incomplete and unreliable view of our customer base.

To resolve this, we grouped 5/6/7 customers as *Priority Access* customers. They are identified based on the use of company email domains, which are more indicative of commercial or institutional buyers. Level 1 will consist of Level 1 and Level 3 customers for simplification when referring to Level 1, which includes all others using personal or generic email providers, such as Gmail, Yahoo, and Outlook.

Connecting Transaction Data to Customers (joining tables)

After left-joining the tables, we further refined and standardized the data. Title case formatting was applied to ensure consistency. To improve clarity, we renamed columns and dropped irrelevant data such as *Product_Name*, as SKUs provide more precise product differentiation. We also reordered the columns to enhance readability and ensure a more logical flow. These refinements ensure a structured, standardized dataset, setting the foundation for deeper insights and predictive analysis.

Variable Dictionary

The table below outlines the variables included in our final, cleaned dataset used for modeling. Each variable is defined with its corresponding description and role in the analysis.

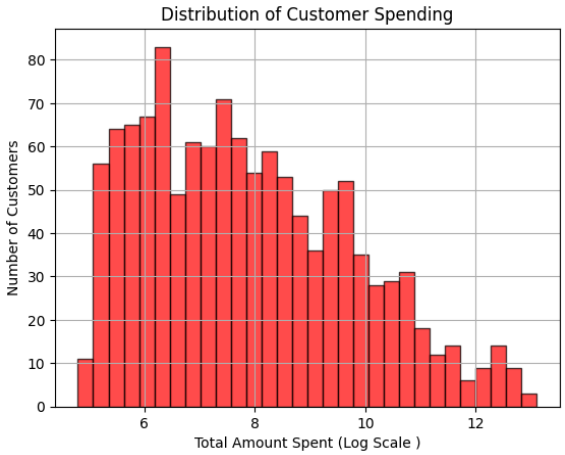
Column Name	Description
Order_Num	Unique identifier for each order.
Order_Date	Date when the order was placed.
Prev_Order_Date	Date of the customer's previous order.
Days_Between_Orders	Number of days between the current order and the previous order.
Customer_ID	Unique identifier for each customer.
First_Name	Customer's first name.
Last_Name	Customer's last name.
Email	Customer's email address.
Email_Domain	The domain is extracted from the customer's email (e.g., gmail.com).
Customer_Type	Binary classification: 0 for personal email, 1 for corporate email.
State	The state associated with the customer's default address.
Zip_Code	The zip code is associated with the customer's default address.
SKU	Stock Keeping Unit – unique product identifier.
Category	Product category classification.
Quantity	Number of units purchased in the order.
Total_Price	Total price for the order (sum of all purchased items).
Total_Amt_Spent	Total amount spent by the customer across all purchases.
Total_Orders	Total number of orders placed by the customer.
Day_of_Month	The day of the month when the order was placed.
Month	The month when the order was placed.
Quarter	The quarter in which the order was placed (Q1, Q2, Q3, Q4).

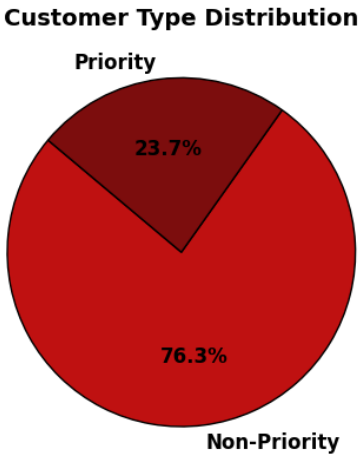
Exploratory Data Analysis

Analysis of the dataset is based on the joined *customers* table and *transaction_details* table, as well as the *products* table. The analysis is divided into customer purchase behavior, product analysis, and time-based trends to gain insight into purchasing patterns and sales performance.

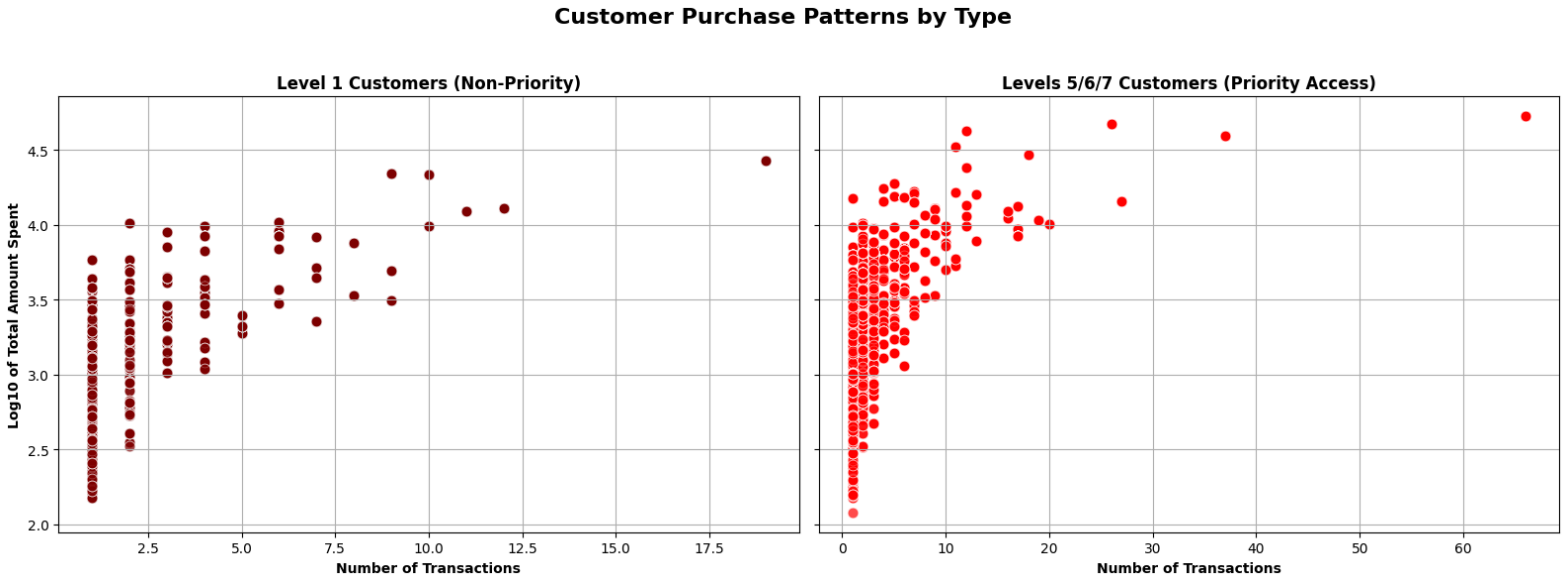
Customer Purchase Behavior

Customer purchase behavior helps the marketing and sales team understand their customers and better segment them for proactive marketing campaigns and sales strategies.





The distribution of customer spending is heavily right-skewed, indicating that most customers spend modest amounts while a small subset contributes disproportionately to total revenue. A substantial portion of customers spend between $400 and $3,000, around $1,000, as seen near the log scale of 6.5–7. Spending drops off sharply after $3,000, but a long tail of high spenders exists, including customers spending over $50,000, although they are rare.

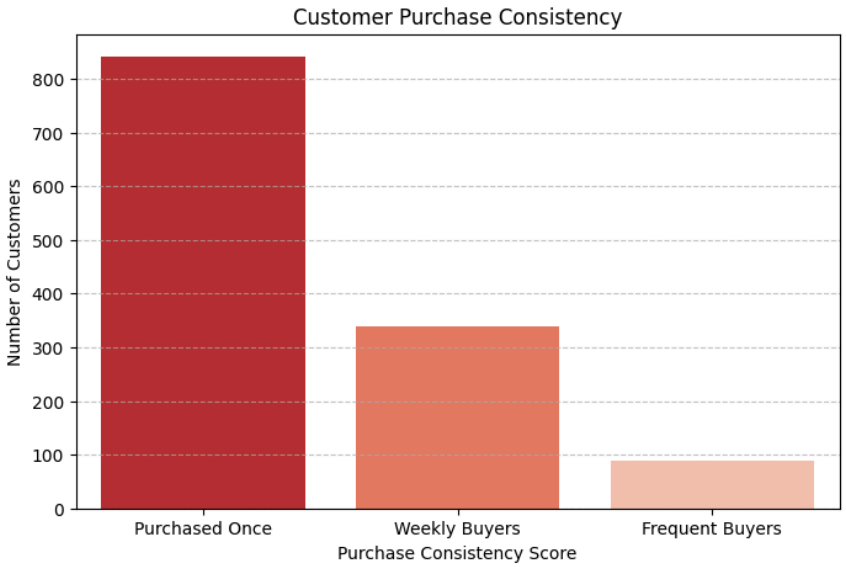


The bulk of the customer base (76.3%) consists of non-priority customers who purchase less frequently and spend less per transaction. Level 1/3 Customers have a transaction count capped at around 10–12, and their spending is relatively uniform, with few exceeding a log10 value of 4.5. This pattern suggests more transactional, one-time behavior and less brand loyalty.

In contrast, Priority Access customers, though only 23.7% of the customer base, account for a significant share of revenue and likely comprise the highest-spending group. They exhibit a much wider range in both spending and frequency, with transaction counts going over 60. While

many still cluster at lower order frequencies, their spending levels are much higher, reflecting greater lifetime value and consistent reordering potential.

Between these two groups, the mid-tier segment (customers spending between $1K and $10K) represents a broad and strategically important cohort. Many of these customers could be nurtured into high-value segments through targeted retention and upselling efforts.

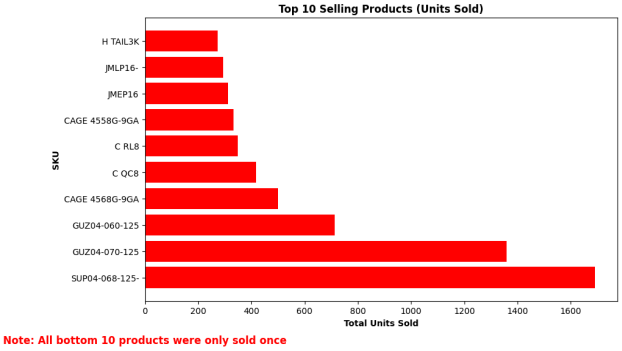


The purchase consistency score is a metric that categorizes customers based on their purchasing frequency, helping uncover behavioral patterns that support targeted marketing and retention strategies. Customers are segmented into three groups based on the number of purchases within a 30-day timeframe:

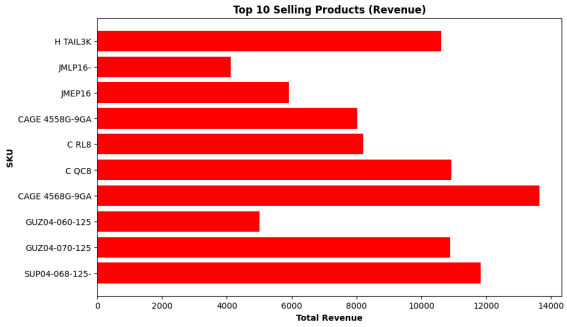
- Purchased Once
- Weekly Buyer (2-4 orders a month)
- Frequent Buyers (multiple times a week, 5 or more orders)

The majority of customers have purchased only once, a smaller segment are weekly buyers, and the smallest group are frequent buyers. This can show low retention and a high churn risk. This means that the few frequent buyers we have are extremely valuable. Strategies can be created to strengthen loyalty and analyzed to mirror these conditions with new buyers.

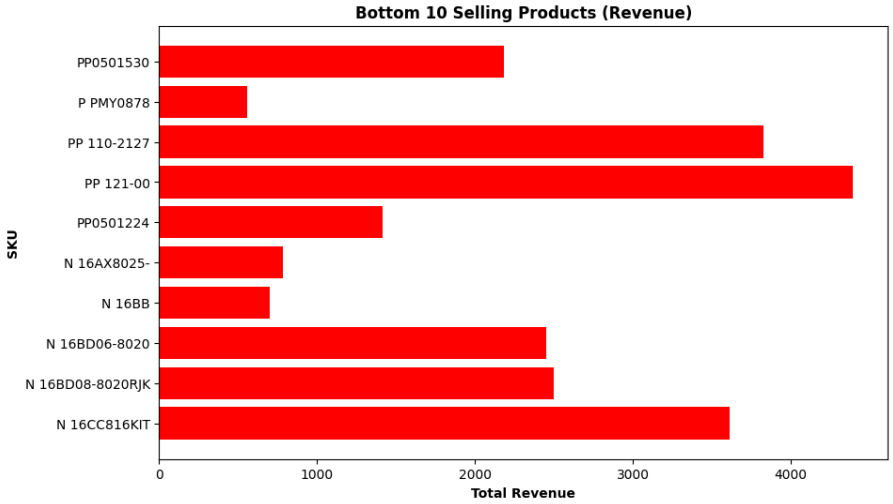
Product Analysis



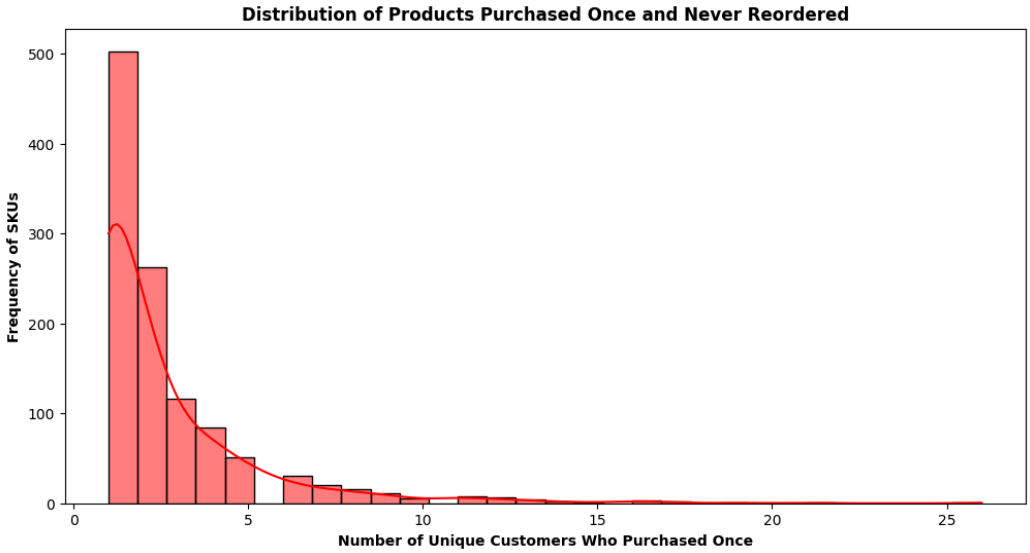
Note: All bottom 10 products were only sold once



Revenue and unit sales do not always align. Some products drive revenue through volume, while others rely on high prices. SUP04-068-125, Super Products Vacuum Truck Filter Bags, has the highest unit sales, but is only mid-ranked in revenue. This is because the variant prices start from $9.74, yet it is heavily relied on in the industry. Piranha hose branded male end, JMLP16- , and mender, JMEP16, swage tools are ranked higher in revenue than in unit sales, but are priced relatively low compared to other products. These are pretty popular products that get sold in a lot of bulk buys and are being sold at mid-volume. This indicates that high-revenue, low-cost products could be potential key revenue drivers.

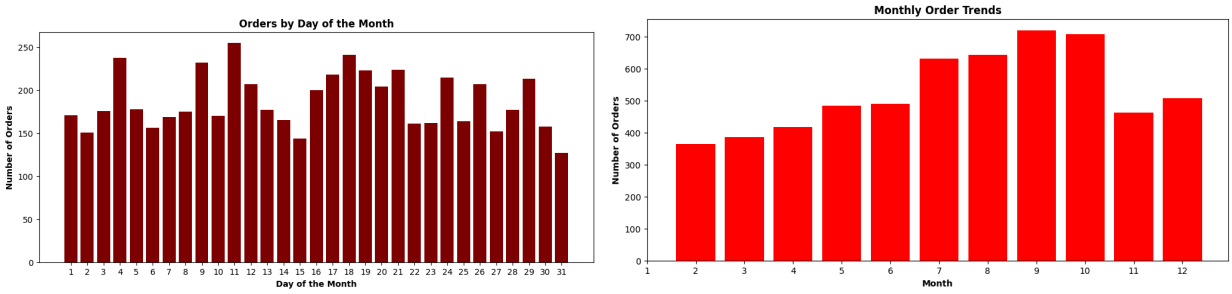


All the bottom 10 products generate less than $25,000 in total revenue and were purchased once. The 1/2" Black Linear Hydro Excavation Gun, PW 300680050BOXOF5-, is an outlier, generating significantly more revenue than the other bottom product. This is because it is a bulk buy item coming in quantities of 5, priced high at $1,224.84. Top revenue generators should be prioritized for stock replenishment. Low-revenue products should be re-evaluated in the sense that they are either niche, overpriced, or have low demand.

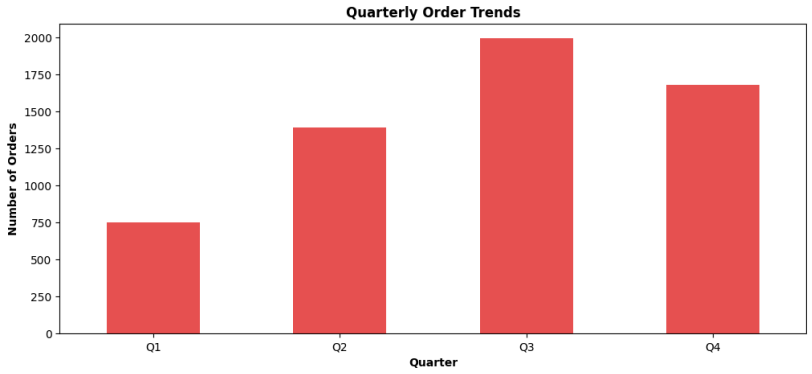


Most SKUs were purchased only once by a single customer. This is because of a wide variety of products offered, where many items are rarely reordered because of site visibility (SEO and UI issues), lack of education, or popularity of the product. Some SKUs have slightly higher one-time purchase rates. These might be niche products or specialized items that attract a few more buyers but still don't see reorders. If many products are one-time purchases, it suggests an opportunity to analyze why customers don't reorder, perhaps due to product satisfaction, lack of awareness, or business model factors.

Time-Based Analysis (Seasonality & Trends, Sales & Inventory Planning)



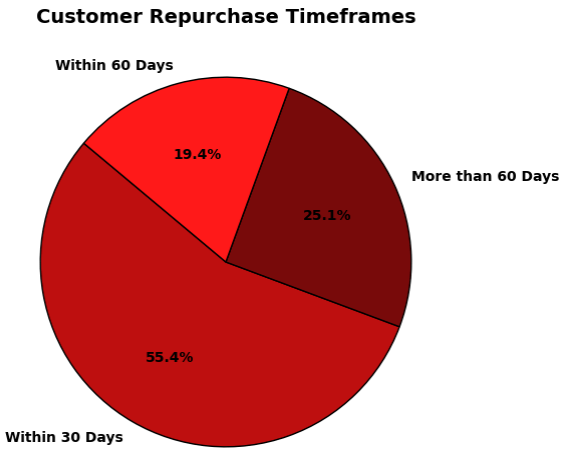
Customers may follow a monthly purchasing cycle, with spikes on predictable days such as the end of the month, and orders, which are slightly elevated, suggesting that some customers finalize purchases before month-end deadlines. Orders show a gradual increase from February through September. June through October maintains strong order volume, with the highest activity in September and October (both exceeding 700 orders). There is a notable drop in November, likely due to seasonal slowdowns or budget constraints. Orders partially recover in December, possibly due to end-of-year purchasing or companies finalizing budgets.



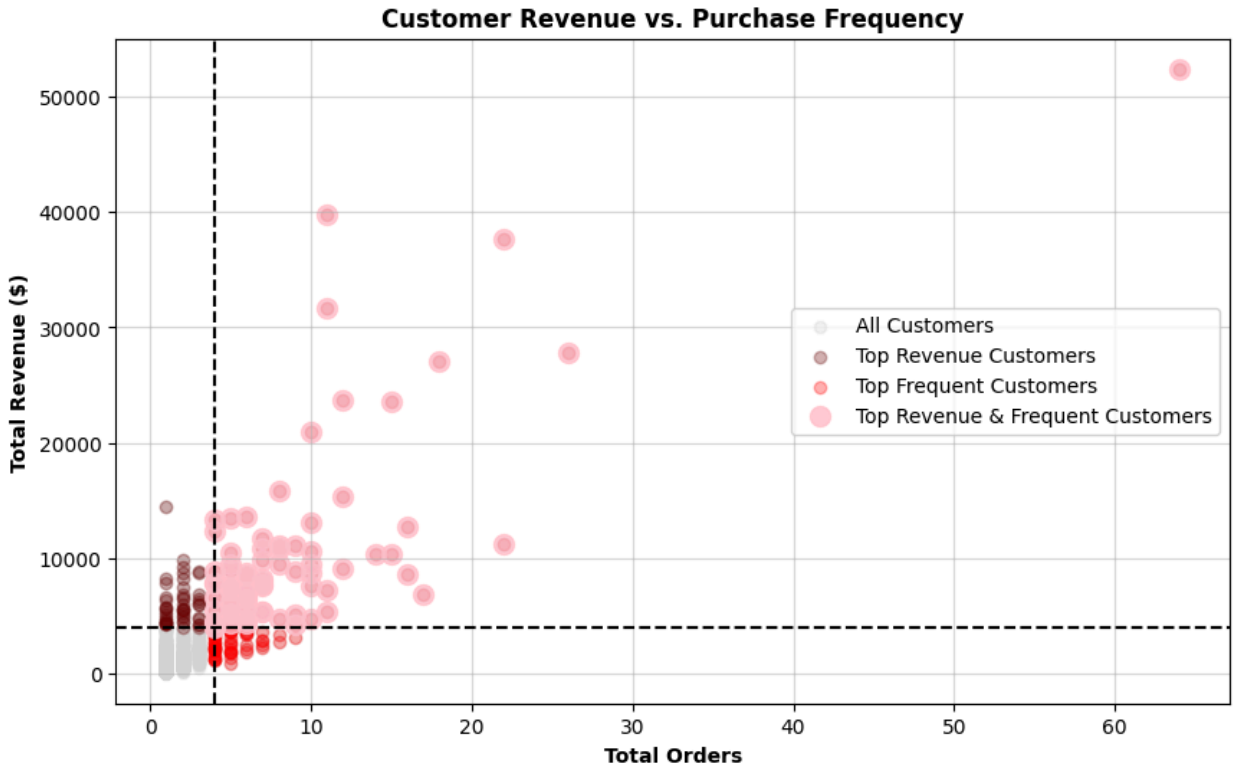
Q1 (Jan-Mar) is the slowest, suggesting a seasonal dip, staying below 800 orders. This is because EPA switched from one website domain to another at the beginning of 2024, with a lack of promotional awareness and SEO strategy to their customers, and had not begun to push for an online initiative. Q3 (Jul-Sep) is the strongest quarter, peaking at 2,000 orders. This is aligned with EPA optimizing its ad spend on Google, increasing traffic to its website.

Customer Repurchase Distribution

A large percentage of customers did not make a second purchase within 30 days. This suggests that many buyers may be one-time purchasers or have longer purchase cycles. The *Days_Between_Orders* metric measures the time between a customer's current purchase date and their last purchase date. To analyze customer retention, we calculated the number of customers who repurchased within different timeframes. Customers were categorized into three groups based on their reorder cycle: those who repurchased within 30 days, indicating frequent buyers with a short buying cycle; those who repurchased between 31 and 60 days, representing moderate repurchasers; and those who placed their next order after 60+ days, highlighting customers with longer buying cycles. This segmentation helps determine the optimal timing for remarketing efforts, identify which customer segments need engagement strategies, improve sales, and customer reorder prediction.



Purchase Frequency vs. High Revenue



Understanding purchasing patterns can be crucial for optimizing retention strategies and customer engagement. In our analysis, we measured *Total Revenue Per Customer* by summing the total purchase amounts for each distinct *Customer_ID*. We also calculated the *Total_Order_Per_Customer* by taking the count of each distinct *Order_Num* placed by each customer. Our findings reveal that high-revenue customers tend to place fewer orders but spend significantly more per transaction. In contrast, frequent buyers place multiple smaller orders, providing consistent revenue over time. Recognizing these differences allows for more effective sales strategies.

Predictive Modeling

Initial Modeling & Key Findings

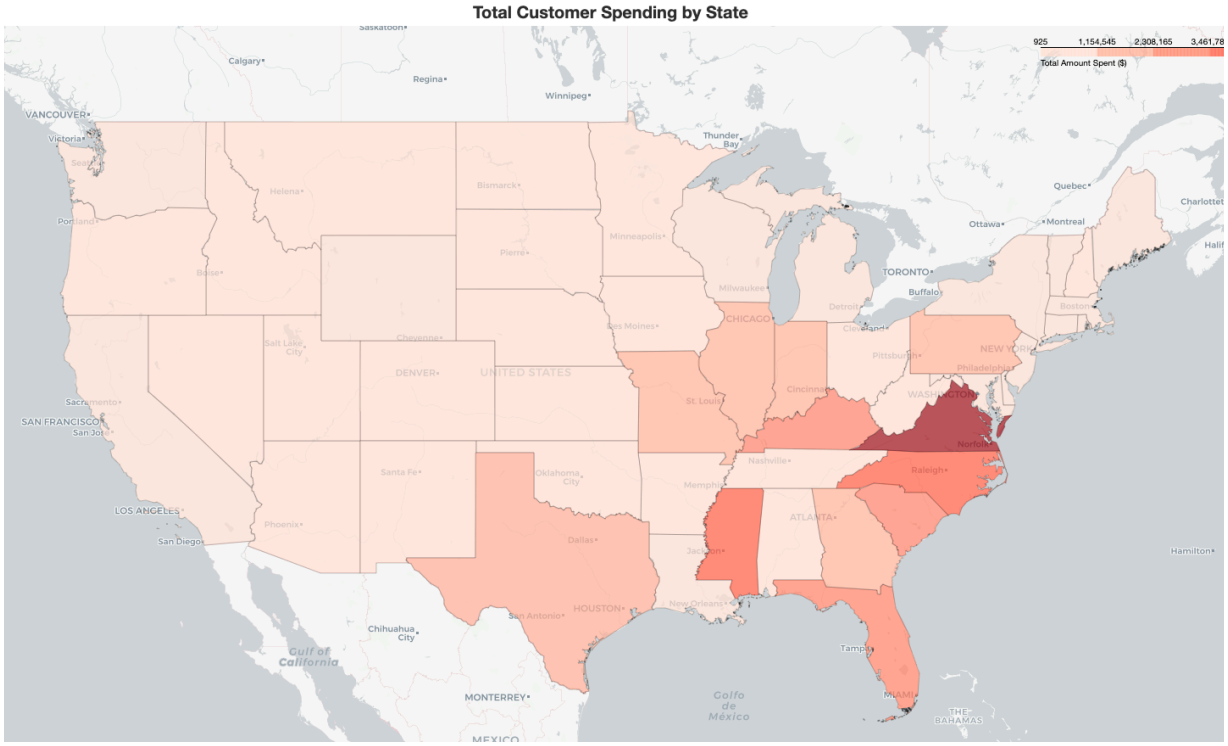
The initial model used K-Means Clustering to group customers based on purchasing behavior, identify patterns, and refine customer segmentation. The Elbow Method was used to determine the optimal number of clusters, which came out to 3 clusters:

Cluster 0: There is an average of 12.6 days between orders with a standard deviation of 35.1. The customers in this cluster are Level 1. These customers tend to buy fewer items per order and spend the least comparatively. This cluster also had a large number from Texas. They

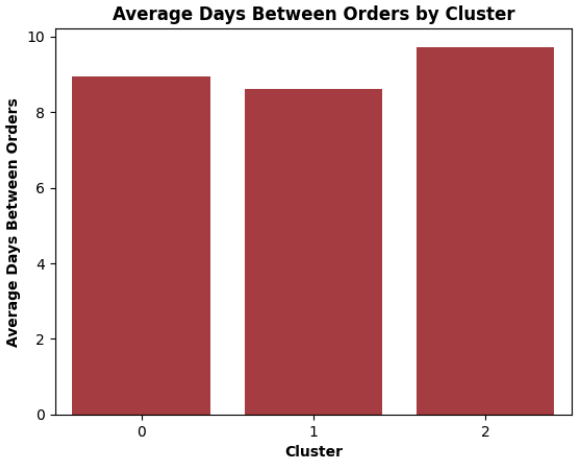
represent new customers and customers who don't purchase all too frequently, as their re-purchasing habits are more spread out.

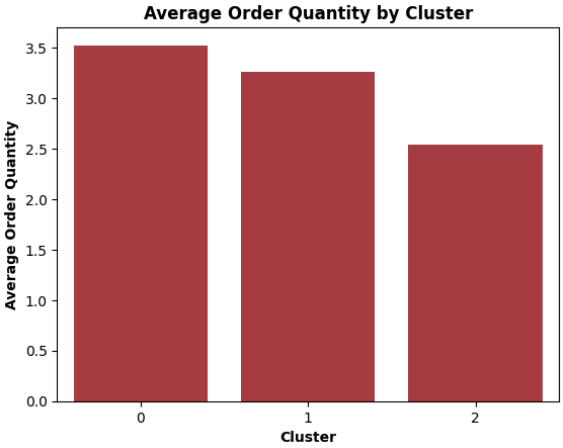
Cluster 1: Average of 11 days between orders with a standard deviation of 30; these customers are level 5/6/7 and have slightly more items per order than cluster 0. Naturally, they spend more per order as well, and many are from Virginia. Cluster 1 consists of returning customers who represent moderately engaged repeat customers.

Cluster 2: Average of 10 days between orders with a standard deviation of 30.6; these customers are also level 5/6/7 and order around 4 items per order, much higher than the previous clusters. They also spend more, and many of them are from Florida. Cluster 2 represents the more frequent buyers who spend the most.

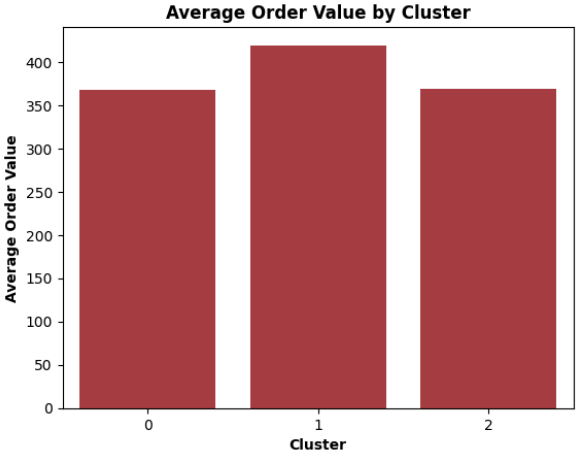


Total revenue generated by customers aggregated by state, the darker the red, the higher the total spend. The top spending states include Virginia, North Carolina, Texas, and Georgia. States like California, New York, and Illinois are not as dark despite population size, suggesting potentially underdeveloped markets.

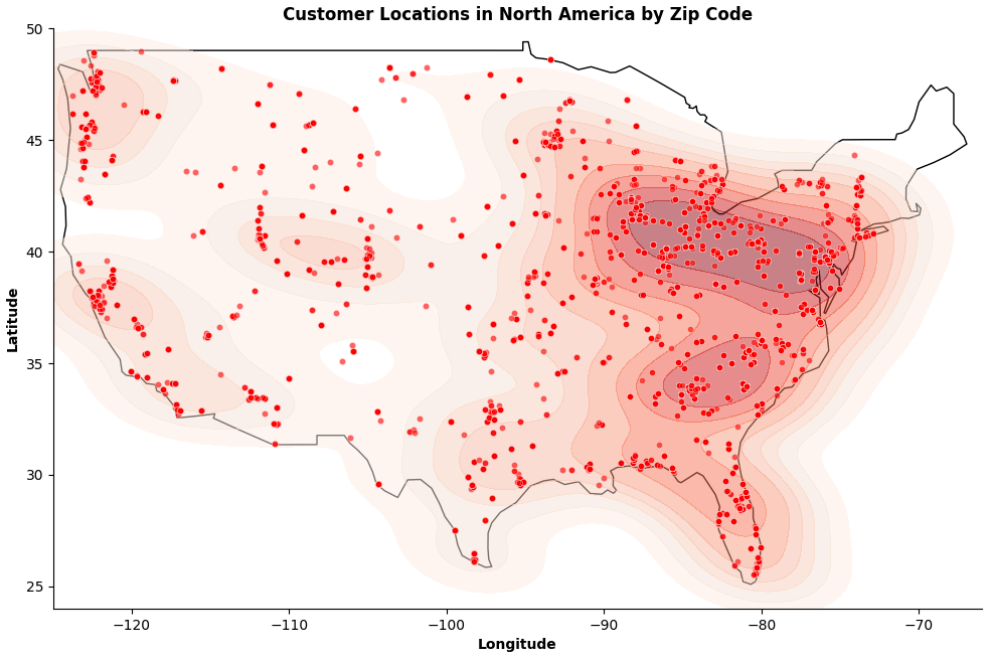




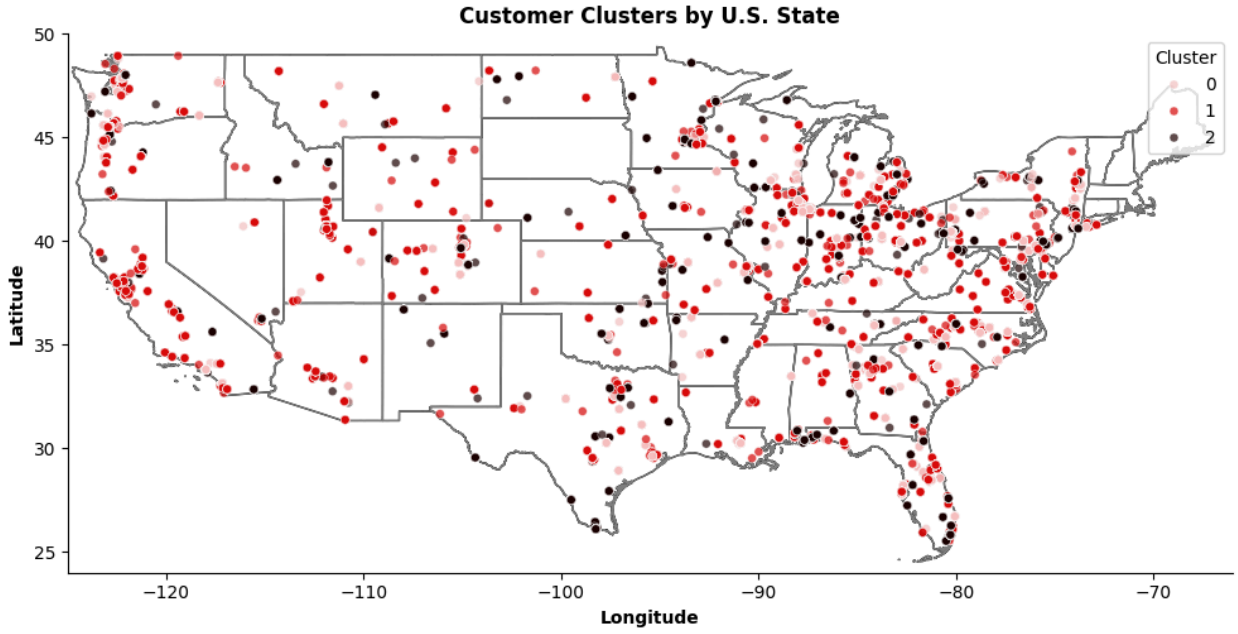
The Average Day between Orders by Cluster shows repurchase churn. Cluster 1 is the most engaged, returning 8.7 days. Cluster 0 follows at 9 days and Cluster 2 is the slowest at under 10 days, making, the least frequent buyer. The Average Order Quantity by Cluster shows that Cluster 0 buys 3.5 items each time, Cluster 1 buys 3.25 items, and Cluster 2 buys 2.5 items per order. This indicates Cluster 2 tends towards smaller baskets.



The Average Order Amount by Cluster shows three distinctive spending tiers, with Cluster 0 spending the most per order, at about $9,800. Cluster 1 spends around $6,800 and Cluster 2 around $5,200. In Average Order Value by Cluster, when you divide spend by quantity, Cluster 1 leads at about $420 per item, while Clusters 0 and 2 are near $370. This indicates that Cluster 1 favors high-priced products.



The darker red the area, the higher the concentration of customers. The Midwest, Southeast, and Northeast are where our customers reside the most. There are sparse but visible clusters on the west coast and southwest regions with pockets of potential.



Cluster 0 (light red) is spread widely across Texas, California, Colorado, and Utah. These are non-priority customers who place fewer but larger orders. Their spatial distribution suggests

regional contractor networks or occasional bulk buyers. These are high-value but under-engaged customers.

Cluster 1 (medium red) has dense pockets in Virginia, North Carolina, and Ohio. Customers in this group are Priority Access, showing high order value and moderate frequency. These customers form your most reliable and profitable base. These customers align with reliable, high-value Priority Access customers.

Cluster 2 (dark red) is heavily concentrated in Florida, Georgia, and the Southeast. These customers order frequently but in smaller quantities. Indicates either just-in-time procurement or smaller-scale businesses with routine purchasing needs. These customers are reliable, high-value Priority Access customers.

Random Forest Model

Goal

The Random Forest Model aimed to identify customers likely to reorder within 30 days of their last purchase, based on each customer's transactional data from 2024. Each customer was assigned a probability rating based on potential reorders within that time window. At a high level, this approach is straightforward. Still, at a more granular level, the model can be used strategically to prioritize outreach, boost retention, and spot potential churn early.

Data Preparation

Before modeling, the data was transformed and cleaned. The original dataset contained Granular transaction data, which was not useful for this particular model; the model only needed 1 record per *Order_Num*. Each of these rows contained dates, monetary value, and customer level; product details were removed because they were not useful in this model. Using *Order_Date,* we were able to create *Days_Between_Orders* and used this column to create a binary column, *Reordered_30_Days,* which allowed us to identify customers who had reordered within 30 days (1) and those who hadn't (0). This was used as our Target variable for the model.

Model Results

The model achieved strong performance metrics across all data splits:

- AUC (Area Under the ROC Curve): ~96.6% across Train, Validation, and Test.
- Misclassification Rate: About 8% across all sets.
- Root Average Squared Error (RASE): Very low and consistent (~0.22 to 0.23) across all sets.
- KS Statistic: ~0.86 to 0.87, indicating excellent separation between reorders and non-reorders.

The model passed all overfitting checks. Train, Validation, and Test performance were aligned, with no major drop-offs between datasets. This suggests the Random Forest model captured underlying patterns in customer purchasing without simply memorizing the training data. Additionally, no data leakage or unrealistic performance (such as perfect classification) was observed.

Important Observations

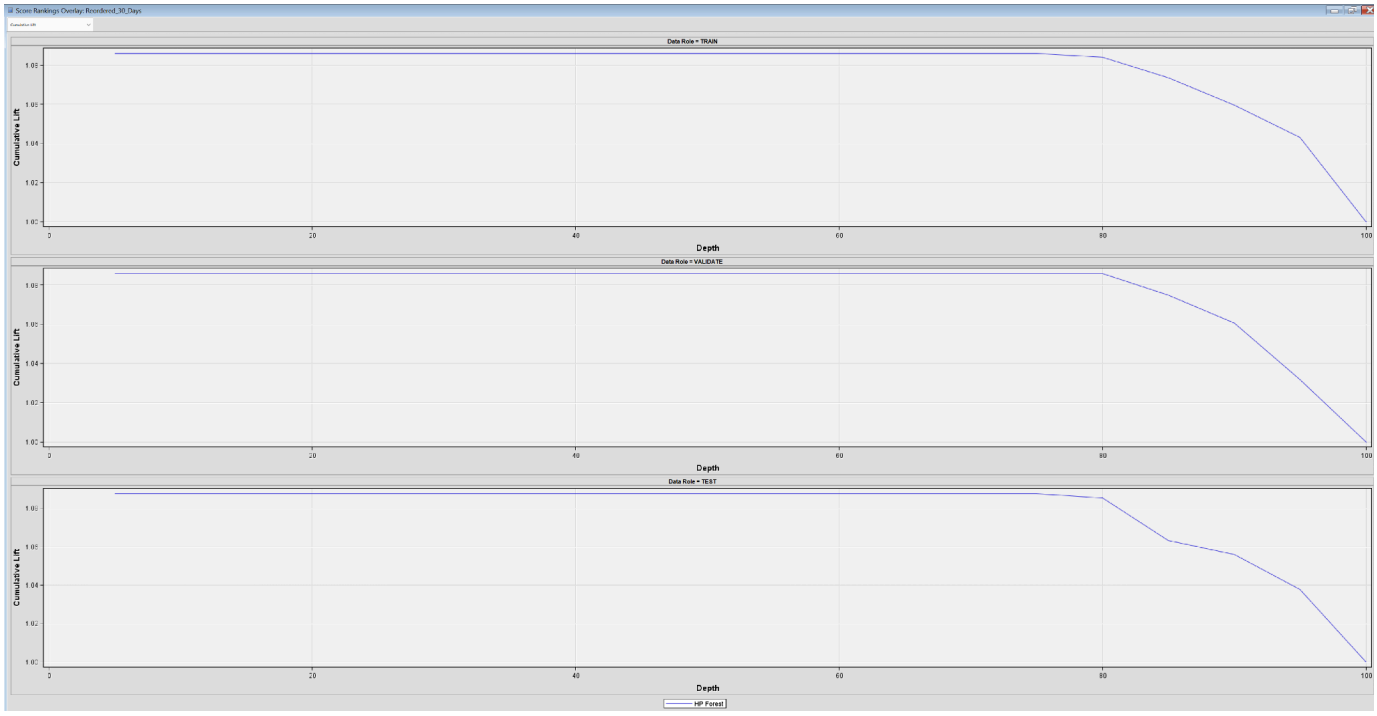
There was no major overfitting: performance on the Validation and Test sets stayed consistent with Train results. The model showed strong generalization ability, meaning it learned real patterns, not just memorized the training data.

Feature Importance

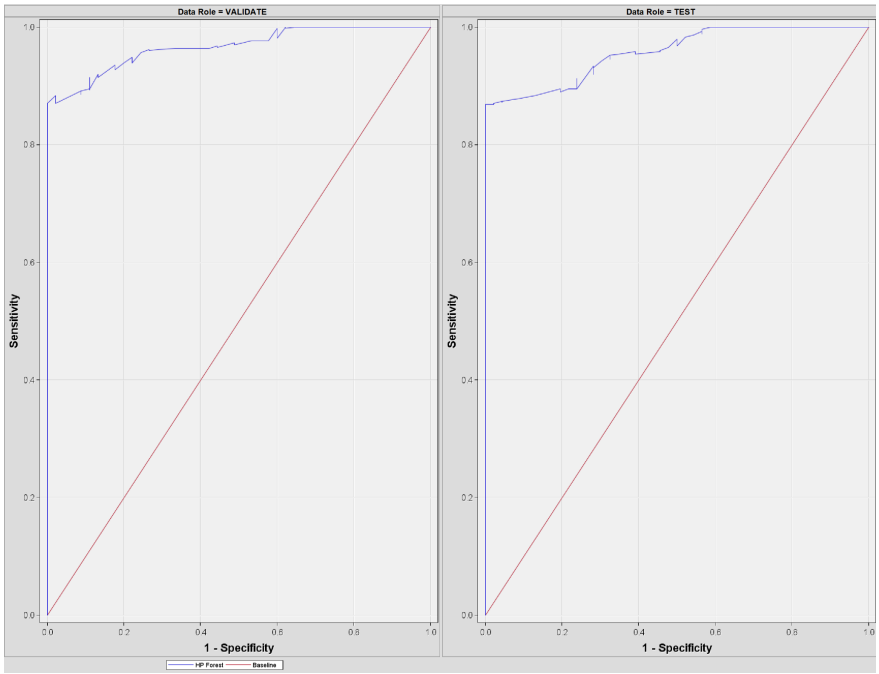
Days_Between_Orders had the largest influence, suggesting that recency patterns are critical for anticipating customer behavior. *Order_Price* contributed meaningfully, indicating that higher transaction values are associated with reorder tendencies. *Customer_Type* also played a role, capturing differences in reorder likelihood based on the level of that customer.
These features make sense logically. Customers who order more frequently and spend more are naturally more likely to reorder.

Visual Validation

Lift Charts show that targeting top-scoring customers would greatly outperform random targeting.



ROC Curves for Train, Validation, and Test sets, all bending sharply toward the top-left corner (high sensitivity and specificity).



Strategic Application

By using this model, the EPA Sales can take a more proactive approach to customer retention. Instead of waiting for customers to drop off, teams can flag high-risk accounts early, those who are unlikely to reorder within 30 days, and follow up with timely interventions like personalized emails, check-in calls, or targeted promotions.

Additionally, the model helps identify the most engaged and profitable buyers, those with a strong likelihood to reorder. These customers are ideal candidates for loyalty-building programs, like personalized discounts, early access to new products, or account-level support.

Because the model performed well on all data splits and didn't show signs of overfitting, it's a dependable tool that can be used in a real-world setting. When integrated into the sales process, this approach can improve customer experience, reduce churn, and help the company focus resources where they matter most.

Association Rules

In the association rule mining analysis, the top ten rules by confidence for the product category were identified. Category was used instead of product to reduce dimensionality, which helped simplify the analysis by decreasing the number of unique items. While the rules show high confidence, indicating a strong likelihood that the consequent occurs given the antecedent, they all exhibit low support. This means that these rules apply to a small part of the dataset. However, the rules have high lift, suggesting the occurrence of the antecedent increases the chance of the

consequent beyond random chance. The combination of high confinement and high lift makes the rules interesting, but the low support limits their generalizability. If the data were expanded over multiple years, then this could potentially help uncover more practical impacts. One interesting finding is that vacuum truck hose fittings are frequently purchased together with other items. This could be a product that complements many others and may represent a strong candidate for cross-selling strategies or bundled promotions.

antecedents	consequents	support	confidence	lift
(vacuum truck tubes, hydro excavation, vacuum ...	(vacuum truck hose fittings)	0.005686	0.666667	2.378744
(vacuum truck tubes, vacuum truck hose)	(vacuum truck hose fittings)	0.005686	0.666667	2.378744
(vacuum truck tubes, hydro excavation)	(vacuum truck hose fittings)	0.015435	0.633333	2.259807
(vacuum truck tubes, vacuum truck hose)	(vacuum truck parts)	0.005280	0.619048	2.801646
(hydro excavation, vacuum truck parts)	(vacuum truck hose fittings)	0.018684	0.589744	2.104274
(vacuum truck valves, vacuum truck parts)	(vacuum truck hose fittings)	0.007311	0.580645	2.071809
(manhole tools, vacuum truck tubes)	(vacuum truck parts)	0.007717	0.575758	2.605726
(hydroexcavation trigger)	(hydro excavation)	0.011779	0.547170	3.111160
(manhole tools, hydro excavation)	(vacuum truck hose fittings)	0.007717	0.527778	1.883172
(jetter hose, vacuum truck tubes)	(vacuum truck hose fittings)	0.007311	0.514286	1.835031

Cox Proportional Hazards Model

The objective of this analysis is to identify the timing of when a customer is most likely to repurchase and to understand the key factors influencing this behavior. We aim to analyze, plot, and rank customers based on their likelihood to repurchase, enabling us to prioritize marketing and sales strategies accordingly. The Cox Proportional Hazards model was selected for this work because it does not assume a specific shape for the baseline time distribution, offering flexibility in modeling diverse customer behaviors. Additionally, it provides interpretable coefficients for covariates, allowing us to clearly understand the impact of different factors on repurchase timing. The model is also well-suited for censored data, which is common when customers have not yet made a repeat purchase within the observation window.

Key factors considered in the model include cumulative spending, representing the total amount a customer has spent previously, and an industry proxy derived from product categories, using domain knowledge to infer customer industry from the types of products purchased. Rather than analyzing purchases at the individual product level, which would introduce excessive complexity, we propose to model behavior based on a limited set of high-level product

categories. This approach enables more interpretable, scalable, and actionable insights for informing business strategies.

Our initial Cox baseline model achieved a concordance of 0.62, a respectable result given our limited number of observations within the data. *Cumulative_Total_Spent* was expected to be significant based on the general assumption in customer lifetime value, retention, and loyalty models. Bigger spenders are usually more loyal, more engaged, and more likely to repurchase faster. Although there is a low p-value of <0.005, showing statistical significance, the coefficient is still 0, showing a neutral effect. While we can statistically detect an association, the practical effect is negligible. In other words, changes in cumulative spending do not meaningfully impact the customer's likelihood to repurchase faster. This indicates that there is no detectable effect on the repurchase hazard, or the “risk” of repurchase. *Prior_Order_Count* has a positive coefficient of 0.03, and exp(coef) is around 1.03, meaning every additional prior order increases repurchase hazard by about 3%. So, customers with more prior orders are more likely to repurchase sooner, specifically, additional prior orders increase the chance the customer will repurchase ~3% faster. *Customer_Type* has a positive coefficient of 0.18 and exp(coef) of 1.19, showing certain customer types are 19% more likely to repurchase faster.

model	lifelines.CoxPHFitter
duration col	'Time_To_Next_Days'
event col	'Repurchased'
baseline estimation	breslow
number of observations	1244
number of events observed	1244
partial log-likelihood	-7428.39
time fit was run	2025-04-27 15:46:23 UTC

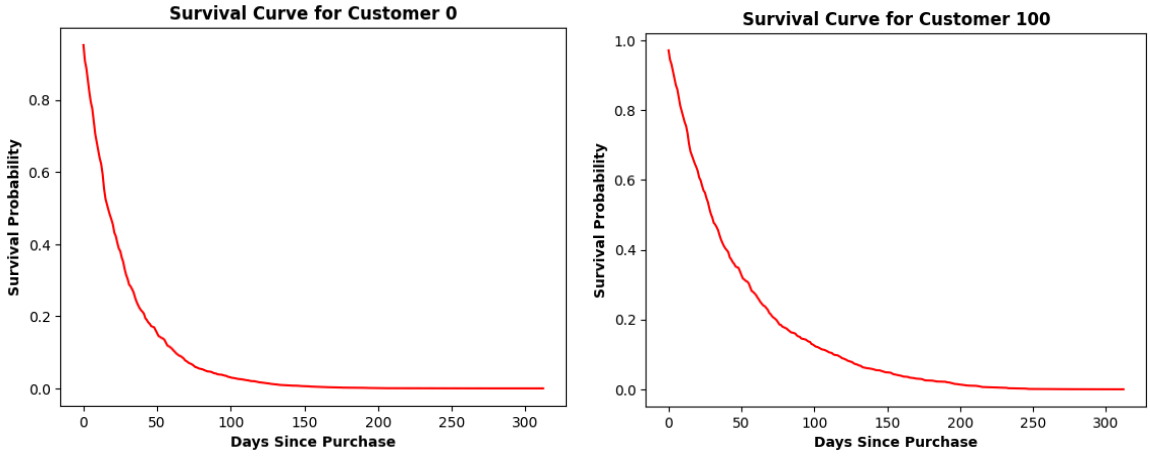
	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	−log2(p)
Cumulative_Spent_Total	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	2.89	<0.005	8.03
Prior_Order_Count	-0.00	1.00	0.01	-0.02	0.01	0.98	1.02	0.00	-0.23	0.82	0.28
Customer_Type	0.11	1.12	0.07	-0.03	0.25	0.97	1.29	0.00	1.58	0.11	3.13
Num_of_Orders	0.03	1.03	0.00	0.03	0.04	1.03	1.04	0.00	13.83	<0.005	142.18
Customer_Tenure_Days	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	0.00	-5.77	<0.005	26.88

Concordance	0.69
Partial AIC	14866.79
log-likelihood ratio test	393.88 on 5 df
-log2(p) of ll-ratio test	273.09

After the initial baseline, we re-ran the model incorporating customer tenure (time since first order) and number of orders to better capture customer lifecycle effects. Adding tenure helps control for the customer's stage in their relationship with the company, ensuring that the model does not confuse "newness" with "low loyalty." Similarly, including the number of orders reflects customer engagement and loyalty, as customers who have purchased multiple times are generally much more likely to return. These additional features provide a more comprehensive and accurate view of customer behavior.

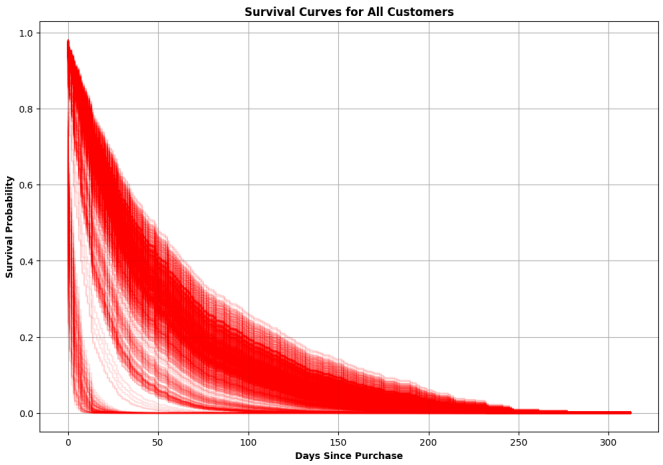
With the addition of customer tenure and number of orders into the model, the results show that *Cumulative_Spent_Total* remains statistically significant but has no practical effect, while *Prior_Order_Count* and *Customer_Type* lose significance after controlling for the additional variables. *Prior_Order_Count* now has a negligible negative coefficient (-0.00) and a high p-value (0.82), indicating no meaningful impact on time to repurchase. This suggests that *Prior_Order_Count* has become redundant, as stronger loyalty indicators are now captured by the number of lifetime orders. Furthermore, this confirms that cumulative spend alone does not drive faster return behavior. For *Customer_Type*, although the coefficient (0.11) implies that certain customer types are associated with a 12% higher likelihood of repurchasing ($\exp(0.11) \approx 1.12$), the evidence is weak ($p = 0.11$). Thus, we cannot confidently conclude that customer type reliably influences repurchase timing based on this dataset.

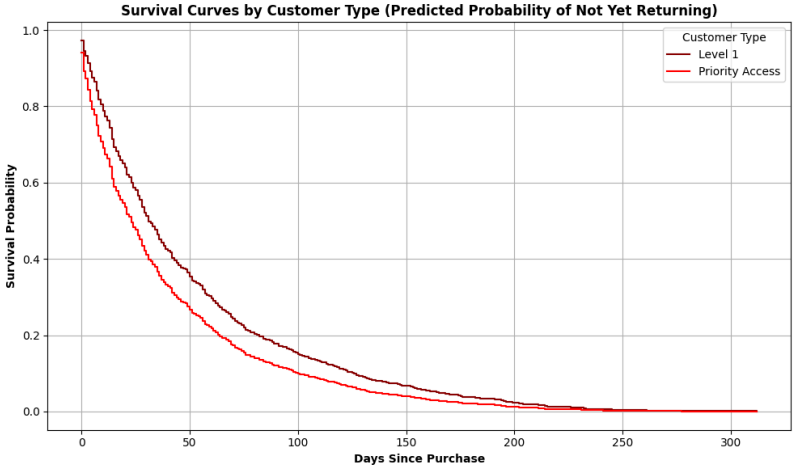
Among the newly added variables, *Num_of_Orders* emerges as a strong predictor. It shows a positive coefficient (0.03) with a highly significant p-value (<0.005), meaning that customers with a higher number of lifetime orders tend to repurchase faster. Specifically, each additional lifetime order increases the hazard of repurchasing by approximately 3%, confirming that prior engagement is a strong driver of loyalty. For *Customer_Tenure_Days*, the coefficient is slightly negative (-0.00) with a statistically significant p-value (<0.005), suggesting that customers with longer tenure are slightly less likely to repurchase quickly. Although the effect size is minimal, this pattern is consistent with the concept of customer fatigue. Operational challenges such as extended turnaround times, long lead times, and poor customer service likely contribute to this fatigue. Customers may become frustrated by delays and rigid company policies, eroding trust and enthusiasm over time, and ultimately slowing their willingness to re-engage with future purchases.



We tested individual customers and their repurchase predictions, customers 0 and 100. Both have a very steep drop-off in survival probability within the first 50 days. Customer 0 is highly likely to repurchase quickly. Most of the repurchase risk (hazard) happens early. If the customer doesn't

repurchase within 100 days, they likely won't come back at all. This indicates that these two customers have similar covariate profiles, and the model is treating them almost identically.





We then analyzed all customers. Almost all curves start near 1.0 with 100% survival, meaning right after purchase, no one has repurchased yet, which makes sense. There is a steep drop early on, meaning a large proportion of customers repurchase quickly, within the first ~50 days. After about 100-150 days, most curves are very close to 0 survival probability, so very few customers are still "active" without having repurchased. There is a thick bundle of curves around 0.2–0.4 probability early (especially between 20–60 days). Some customers drop very fast (early churners, steep drop), and others linger longer; their survival curve is more gradual. Customers are not identical. Some are more "sticky" (survive longer before leaving), while others churn fast. EPA Sales has a "short repurchase window". Customers decide fairly quickly whether to repurchase. After ~100 days, the repurchase probability is extremely low across the board, likely indicating the need for early lifecycle marketing; waiting too long will result in losing customers.

For more meaningful analysis, we took a look at the survival curve by customer type. Priority Access customers are more likely to come back sooner because their survival probability falls faster. Level 1 customers take longer to repurchase. They "survive" (haven't repurchased) for longer compared to Priority Access. The gap is visible especially between day 0 to around day 100, after which both groups stabilize.

Results & Recommendations

Challenges & Adjustments

One of the biggest challenges faced was managing data constraints and cleanup. Initially, customer tags were planned to identify Levels 1, 5, 6, and 7. However, the column was largely incomplete, with mostly null values, since tags were only assigned to customers with created accounts, excluding a significant portion of transaction data. To address this, email domains were used to segment customers into Priority Access (Levels 5/6/7) and Level 1. This approach was based on the high likelihood that customers using company domain emails belonged to Priority Access, while others fell into Level 1. Additionally, aggregation and granularity posed challenges during data cleaning and table joins due to the dataset's initial disorganization. Issues were often identified later in the exploratory data analysis (EDA) process, requiring adjustments and re-cleaning to ensure data integrity.

Comparing Models

The Random Forest model was focused on predicting if a customer will reorder within 30 days? It performed exceptionally well, achieving an AUC of 96.6% and a low misclassification rate of 8%. It was able to effectively identify high-probability reorders and gave us insight into which features mattered most like Days Between Orders, Total Price, and Customer Type. These results give the company a reliable tool for flagging customers at high risk of churn and surfacing those most likely to reorder soon.

Meanwhile, the Cox Proportional Hazards model tackled a different but equally important angle, when customers are likely to repurchase. It unveiled the majority of customers who *do* reorder, do so within the first 50 to 100 days. The model highlighted that customer tenure and number of prior orders were strong indicators of repurchase timing. This model is especially useful for planning follow-ups and designing time-based campaigns.

K-Means Clustering gave us a much clearer picture of our different types of customers. We were able to separate them into three distinct segments based on spend and frequency, helping us tailor future marketing and product strategies to fit each group’s behavior. It also pointed out opportunities in geographic targeting and customer development, especially in mid-tier and under-penetrated states.

Finally, Association Rule Mining uncovered products that are commonly purchased together. While the purchasing patterns we found applied to a small subset of transactions, they still give us ideas for cross-sell opportunities and bundling strategies, especially for things like hose fittings and accessories that often go hand-in-hand.

Overall, the models we used are each useful in their own way and when used together they create a good system for predictive analysis. Random Forest helps with targeting, Cox helps with timing, clustering helps with segmentation, and association rules help with product pairing. When combined, they form a full-circle view of the customer and open the door for a much smarter, more data-driven approach to sales and retention.

Results & Recommendations

This project looks at predicting if and when a customer will repurchase and breaks down their purchase behaviors into clusters. After conducting extensive exploratory data analysis and evaluating multiple models, we propose five strategic pillars to guide EPA Sales' marketing, sales, and operations initiatives: customer segmentation, targeted marketing with CRM integration, inventory and promotional optimization, operational efficiency, and loyalty-driven upsell programs.

Segmentation

- The Cox Model shows Priority Access customers repurchase faster than Level 1 customers. We can use *Num_Orders* and *Customer_Tenure_Days* to further refine our customer segmentation. The Cox Extended Model showed that *Num_Orders* and *Customer_Tenure_Days* were the most impactful in time prediction. With this valuable insight, we can implement marketing strategies to support this.
- Nurture mid-tier segments into high-value customers. Mid-tier segments purchase between $1K-$10K.
- Enhancing customer profiles through better checkout tagging and firmographic enrichment will further refine segmentation and ensure future marketing is laser-focused.

Marketing Targeting & CRM Integration

- Geographical heatmaps reveal under-penetrated opportunities in California, New York, and Illinois; shifting a portion of digital ad spend and outbound sales efforts to these states, paired with localized case studies or customer testimonials, can help EPA Sales capture untapped market share.
- By integrating our Random Forest and Cox model scores into the CRM, sales teams can flag high-churn-risk customers, those whose survival probability drops sharply after 50 days, for proactive outreach, and identify top-decile leads for VIP treatment, such as dedicated account managers or early product previews.

Inventory, Promotions, & Operations

- Align restocking notifications and discounts to the identified repurchase window for customers segmented further from the Cox model using *Num_Orders* and

Customer_Tenure_Days. Our survival analysis shows that most repurchases occur within the first 100 days, so implementing automated touchpoints like personalized reminders, product usage tips, or special offers at 30, 60, and 90 days post-purchase can dramatically improve retention and reduce churn risk.

- A large number of SKUs see only one purchase due to site visibility challenges and a lack of customer education, so enhancing SEO, adding "frequently repurchased with…" recommendations, and producing quick-start guides or short product videos will help customers discover complementary items and understand niche products, driving more repeat orders.
- Inventory and promotion strategies should focus on high-volume and high-revenue items, such as the SUP04-068-125 filter bags and popular Piranha hose ends and swage tools, while avoiding overuse of flash sales that risk training buyers to wait for discounts, thereby preserving margins and demand stability.
- Leverage association-rule insights to bundle frequently co-purchased items (e.g., hose fittings with complementary accessories) in promotions and on-site "Frequently Bought Together" recommendations to drive cross-sell.

Loyalty & Upsell Programs

- Mid-tier segments should have tailored upsell campaigns and volume-discount offers to help shift engagement towards stronger loyalty.
- Launching a tiered loyalty program that rewards weekly and frequent buyers with perks like free shipping, surprise product samples, or early access to new SKUs will foster deeper engagement and mirror the "stickiness" of Priority Access customers.

Timing

- Concentrate outreach within the first 50 days post-purchase because beyond ~100 days, conversion chances drop sharply. There is an early repurchase window where most customers who purchase within 50 days will reorder. By 100-150 days, survival curves approach 0 for the majority, meaning they stop purchasing altogether after that. There is a cluster of curves that are considered to have variability, with some being high-churn customers and others "surviving" longer between 20-60 days.

By aligning "who" (clustering), "which" (Random Forest), "what" (association rules), and "when" (survival analysis), EPA Sales can execute a fully integrated, data-driven strategy that boosts customer retention, cross-sell revenue, and operational efficiency.

Works Cited

1. De, T. S., Singh, P., & Patel, A. (2024). A Machine learning and Empirical Bayesian Approach for Predictive Buying in B2B E-commerce. *The 8th International Conference on Machine Learning and Soft Computing (ICMLSC 2024)*, 17–24. https://doi.org/10.1145/3647750.3647754

2. Emre Deniz, E., & Bülbül, S. Ç. ˘Glu. (2024). "Customer Purchase Prediction Using Supervised Machine Learning." ADBA Information Technology and Publishing Limited Company, 1(1). https://doi.org/10.69882/adba.iteb.2024071

3. Jia, Y., & Wang, J. (2019). Customer revisit prediction using Cox proportional hazard model with deep learning. *IEEE Access, 7*, 39452–39461. https://doi.org/10.1109/ACCESS.2019.2906433