customer_360_clean

August 21, 2025

1 Customer 360 - Project Context

This Customer-360 extends our earlier work—same cleaned data foundation, now focused on retention & personalization to fix the leaky bucket.

Findings from previous analysis: - Stagnation (–1.2% YoY in 2011) - 37% new vs 37% lost customers - Rvenue dependence on loyalists (~44% of base → ~88% of revenue)

The focus is on: - Identifying which customers are **most valuable** and protecting them. - Predicting which customers are **at risk of churn** and intervening early. - Designing a **personalized recommendation strategy** to drive revenue, cross-sell for stickiness, and reactivate churned customers.

We address these questions in four phases:

1. **Segmentation & Impact Analysis**
    - RFM-style segmentation (Champions, Loyal, Potential Loyalists, At Risk, Lost).
    - Revenue and activity contribution per segment.
2. **Churn Prediction**
    - Machine learning models (Logistic Regression, Random Forest, XGBoost).
    - Identify customers with high churn probability within a 90-day window.
3. **Product Stickiness & Retention Analysis**
    - Category-level retention, cross-sell patterns, and CLV impact.
    - Identify categories that act as **revenue drivers** vs **retention anchors**.
4. **Recommendation System**
    - Three personalized scenarios:
        - Drive Revenue: recommend top-selling items in favorite category.
        - Cross-Sell Stickiness: recommend products from related categories.
        - Churn Reactivation: recommend sticky-category products with offers.

2 Project Setup

```
All libraries imported successfully!
```

```
Dataset shape: (757349, 15)
Date range: 2009-12-01 07:45:00 to 2011-12-09 12:50:00
Number of unique customers: 5819
```

```
Number of unique products: 4604
Total revenue: $14,891,236.64

   order_id product_id                  product_description  quantity  \
0    489434      85048  15CM CHRISTMAS GLASS BALL 20 LIGHTS        12
1    489434     79323P                   PINK CHERRY LIGHTS        12
2    489434     79323W                  WHITE CHERRY LIGHTS        12
3    489434      22041         RECORD FRAME 7" SINGLE SIZE         48
4    489434      21232       STRAWBERRY CERAMIC TRINKET BOX        24

           order_date  unit_price  customer_id         country  total_amount  \
0  2009-12-01 07:45:00        6.95      13085.0  United Kingdom          83.4
1  2009-12-01 07:45:00        6.75      13085.0  United Kingdom          81.0
2  2009-12-01 07:45:00        6.75      13085.0  United Kingdom          81.0
3  2009-12-01 07:45:00        2.10      13085.0  United Kingdom         100.8
4  2009-12-01 07:45:00        1.25      13085.0  United Kingdom          30.0

   year  month  quarter day_of_week month_year   product_category
0  2009     12        4     Tuesday    2009-12  CHRISTMAS_HOLIDAY
1  2009     12        4     Tuesday    2009-12    BEAUTY_PERSONAL
2  2009     12        4     Tuesday    2009-12    BEAUTY_PERSONAL
3  2009     12        4     Tuesday    2009-12         HOME_DECOR
4  2009     12        4     Tuesday    2009-12  FURNITURE_STORAGE
```

We have cleaned the dataset in the separate analysis, so we expect this dataset doesn't need extensive preprocessing anymore

```
1. Missing Values:
order_id               0
product_id             0
product_description    0
quantity               0
order_date             0
unit_price             0
customer_id            0
country                0
total_amount           0
year                   0
month                  0
quarter                0
day_of_week            0
month_year             0
product_category       0
dtype: int64

2. Data Types:
order_id                 int64
product_id              object
```

```
product_description     object
quantity                 int64
order_date              object
unit_price             float64
customer_id            float64
country                 object
total_amount           float64
year                     int64
month                    int64
quarter                  int64
day_of_week             object
month_year              object
product_category        object
dtype: object

3. Duplicate Records:
Total duplicates: 24766

4. Customer ID Analysis:
Missing customer IDs: 0
Customer ID data type: float64

5. Data Anomalies:
Negative quantities: 0
Negative unit prices: 0
Negative total amounts: 0

6. Basic Statistics:

            order_id       quantity     unit_price   customer_id  \
count  757349.000000  757349.000000  757349.000000  757349.00000
mean   537562.650369      12.337442       2.861636   15347.90287
std     26713.192113      70.291119       3.927546    1692.70848
min    489434.000000       1.000000       0.030000   12346.00000
25%    515100.000000       2.000000       1.250000   13999.00000
50%    537050.000000       5.000000       1.950000   15301.00000
75%    561894.000000      12.000000       3.750000   16814.00000
max    581587.000000   19152.000000     295.000000   18287.00000

        total_amount           year          month        quarter
count  757349.000000  757349.000000  757349.000000  757349.000000
mean       19.662318    2010.424951       7.528010       2.828522
std        60.383443       0.565893       3.443772       1.133337
min         0.060000    2009.000000       1.000000       1.000000
25%         4.350000    2010.000000       5.000000       2.000000
50%        10.500000    2010.000000       8.000000       3.000000
75%        17.850000    2011.000000      11.000000       4.000000
```

```
max      8925.000000    2011.000000      12.000000       4.000000
```

3 Phase 1: Creating a single customer view aggregation

In this phase, we will create an aggregated table in customer level for our transaction data, the goal is to get the clear metric that reflect the behavior and quality of our customer.

```
=== DATA PREPROCESSING ===
Rows before removing missing customer_ids: 757349
Rows after removing missing customer_ids: 757349

Final dataset shape: (757349, 18)
Date range: 2009-12-01 07:45:00 to 2011-12-09 12:50:00
Analysis period: 738 days
```

```
CUSTOMER LEVEL AGGREGATIONS
```

```
Customer base size: 5819
Reference date for recency calculation: 2011-12-09 12:50:00
```

```
   customer_id  total_orders      first_purchase       last_purchase  \
0      12346.0             2 2010-03-02 13:08:00 2010-06-28 13:53:00
1      12347.0             8 2010-10-31 14:20:00 2011-12-07 15:52:00
2      12348.0             5 2010-09-27 14:59:00 2011-09-25 13:13:00
3      12349.0             3 2010-04-29 13:20:00 2011-11-21 09:51:00
4      12350.0             1 2011-02-02 16:01:00 2011-02-02 16:01:00
5      12351.0             1 2010-11-29 15:23:00 2010-11-29 15:23:00
6      12352.0             7 2010-11-12 10:20:00 2011-11-03 14:37:00
7      12353.0             2 2010-10-27 12:44:00 2011-05-19 17:47:00
8      12354.0             1 2011-04-21 13:11:00 2011-04-21 13:11:00
9      12355.0             2 2010-05-21 11:59:00 2011-05-09 13:49:00

   total_spent  avg_order_value  total_transactions  total_quantity  \
0       169.36         7.056667                  24              24
1      5633.32        22.266087                 253            3286
2      1658.40        36.052174                  46            2704
3      3405.99        20.895644                 163            1435
4       294.40        18.400000                  16             196
5       300.93        14.330000                  21             261
6      1459.18        18.470633                  79             570
7       406.76        16.948333                  24             212
8      1079.40        18.610345                  58             530
9       947.61        27.074571                  35             543

   unique_products  days_since_first_purchase  days_since_last_purchase  \
0               24                        646                       528
1              126                        403                         1
2               24                        437                        74
```

```
3            133                           588                        18
4             16                           309                       309
5             21                           374                       374
6             61                           392                        35
7             23                           408                       203
8             58                           231                       231
9             35                           567                       213

   customer_lifespan_days  purchase_frequency
0                     118            6.134454
1                     402            7.245658
2                     362            5.027548
3                     570            1.917688
4                       0          365.000000
5                       0          365.000000
6                     356            7.156863
7                     204            3.560976
8                       0          365.000000
9                     353            2.062147
```

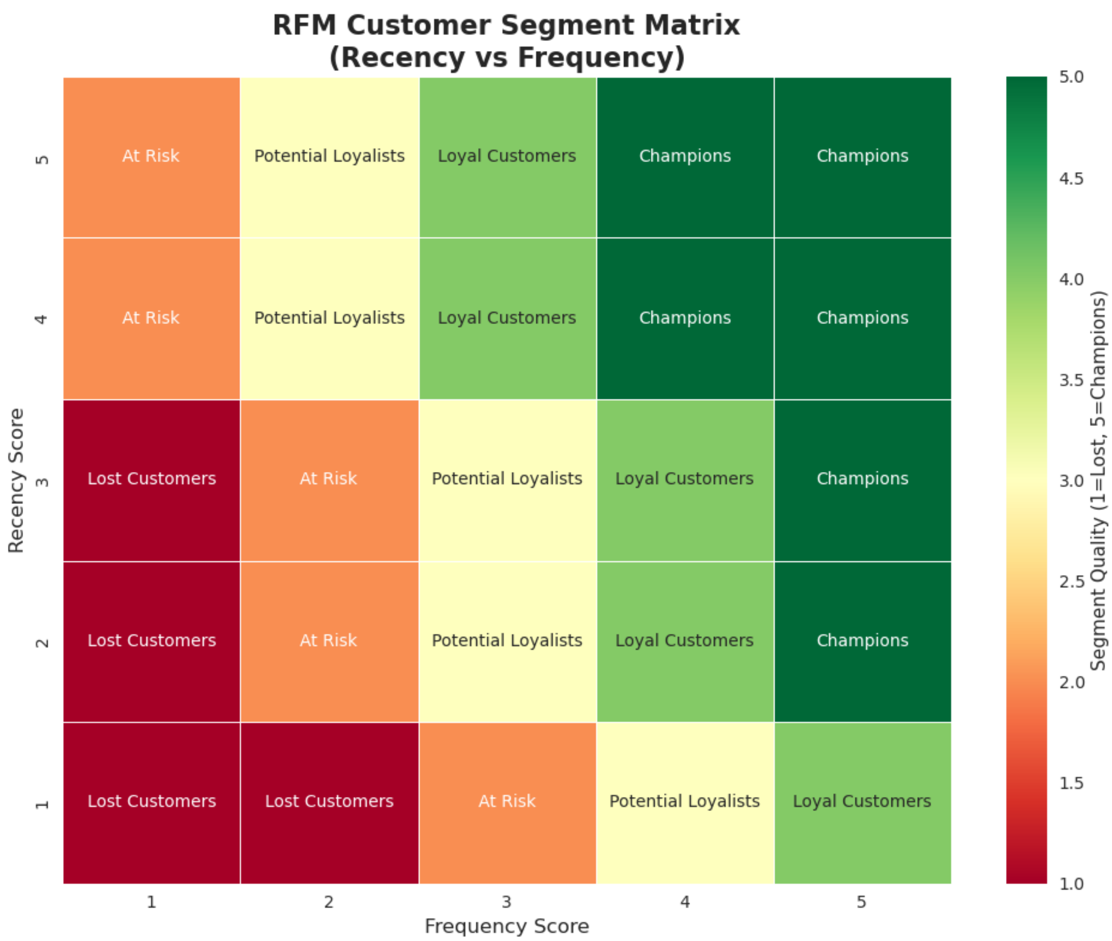
Now we have the aggregated table for our customer, next we will do the RFM (Recency Frequancy Monetary) analysis to segment our customer based on their quality

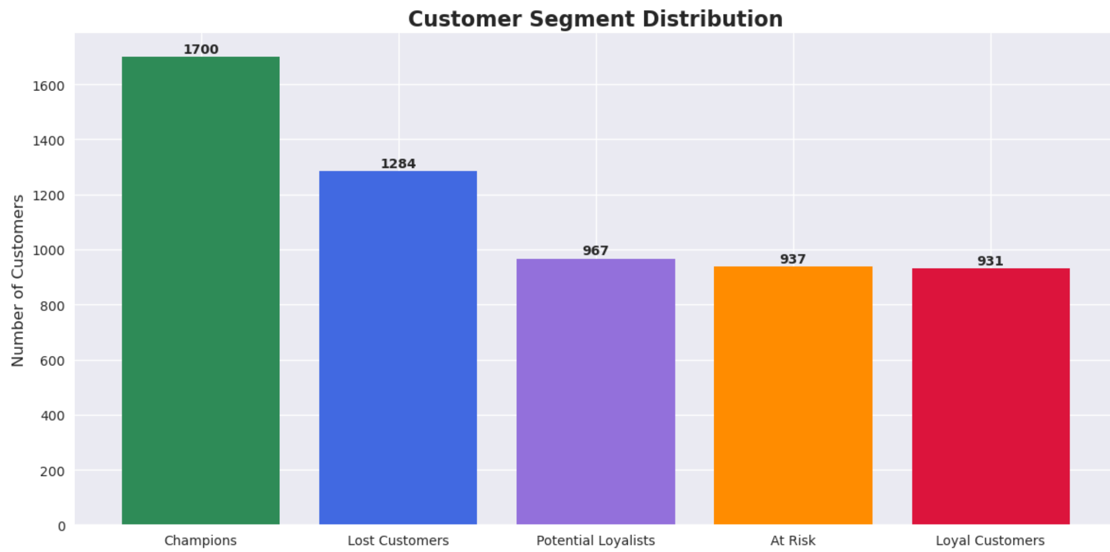
```
RFM ANALYSIS
RFM Score Distribution:
customer_segment
Champions              1700
Lost Customers         1284
Potential Loyalists     967
At Risk                 937
Loyal Customers         931
Name: count, dtype: int64

   customer_id  recency  frequency  monetary r_score f_score m_score  \
0      12346.0      528          2    169.36       1       2       1
1      12347.0        1          8   5633.32       5       4       5
2      12348.0       74          5   1658.40       3       4       4
3      12349.0       18          3   3405.99       5       3       5
4      12350.0      309          1    294.40       2       1       2
5      12351.0      374          1    300.93       2       1       2
6      12352.0       35          7   1459.18       4       4       4
7      12353.0      203          2    406.76       2       2       2
8      12354.0      231          1   1079.40       2       1       3
9      12355.0      213          2    947.61       2       2       3

  rfm_score  rfm_score_numeric customer_segment
0       121                  4   Lost Customers
1       545                 14        Champions
```

2	344	11	Loyal Customers
3	535	13	Champions
4	212	5	Lost Customers
5	212	5	Lost Customers
6	444	12	Champions
7	222	6	At Risk
8	213	6	At Risk
9	223	7	At Risk





We have segmented our customer, this reflect a customer quality progression where

1. "Lost Customer" indicates Churned customer
2. "At Risk" have high risk turned into Churned customer in next few months
3. "Potential Customer" their behavior indicates they can be converted into loyal and repeating customer
4. "Loyal Customer" our frequent and repeating customer
5. "Champions" our loyal customer that drives big portion of our revenue

We will use this segmentation for our Customer 360 table

```
CUSTOMER 360 DATASET CREATION
Final Customer 360 dataset shape: (5819, 22)
Features included: ['customer_id', 'total_orders', 'first_purchase',
'last_purchase', 'total_spent', 'avg_order_value', 'total_transactions',
'total_quantity', 'unique_products', 'days_since_first_purchase',
'days_since_last_purchase', 'customer_lifespan_days', 'purchase_frequency',
'r_score', 'f_score', 'm_score', 'rfm_score', 'rfm_score_numeric',
'customer_segment', 'is_active', 'customer_tenure_months',
'avg_products_per_order']

Customer Segment Distribution:
                     Customer_Count  Avg_Revenue  Total_Revenue  Avg_Orders  \
customer_segment
At Risk                         925       517.08      478300.44        1.78
Champions                      1703      6759.63    11511655.36       14.81
Lost Customers                 1291       237.69      306861.37        1.14
Loyal Customers                 931      1517.45     1412747.39        4.65
Potential Loyalists             969       893.80      866092.69        2.93
```

```
                     Avg_Days_Since_Last_Purchase
customer_segment
At Risk                                    257.02
Champions                                   35.18
Lost Customers                             456.06
Loyal Customers                            109.49
Potential Loyalists                        179.77

Sample of Customer 360 Dataset:

   customer_id  total_orders      first_purchase       last_purchase  \
0      12346.0             2 2010-03-02 13:08:00 2010-06-28 13:53:00
1      12347.0             8 2010-10-31 14:20:00 2011-12-07 15:52:00
2      12348.0             5 2010-09-27 14:59:00 2011-09-25 13:13:00
3      12349.0             3 2010-04-29 13:20:00 2011-11-21 09:51:00
4      12350.0             1 2011-02-02 16:01:00 2011-02-02 16:01:00

   total_spent  avg_order_value  total_transactions  total_quantity  \
0       169.36         7.056667                  24              24
1      4921.53        22.169054                 222            2967
2      1658.40        36.052174                  46            2704
3      3405.99        20.895644                 163            1435
4       294.40        18.400000                  16             196

   unique_products  days_since_first_purchase  …  purchase_frequency  \
0               24                        646  …            6.134454
1              126                        403  …            7.245658
2               24                        437  …            5.027548
3              133                        588  …            1.917688
4               16                        309  …          365.000000

   r_score  f_score  m_score  rfm_score  rfm_score_numeric customer_segment  \
0        1        2        1        121                  4   Lost Customers
1        5        4        5        545                 14        Champions
2        3        4        4        344                 11  Loyal Customers
3        5        3        5        535                 13        Champions
4        2        1        2        212                  5   Lost Customers

   is_active customer_tenure_months  avg_products_per_order
0      False              21.222076               12.000000
1       True              13.239159               15.750000
2       True              14.356110                4.800000
3       True              19.316689               44.333333
4      False              10.151117               16.000000

[5 rows x 22 columns]
```

4 Phase 2: Customer Segment Deep Dive

Customer segmentation from the RFM analysis is important for us to breakdown the main driver on why customer become churned and how they end up spend more in our product. In this analysis we will characterize each segment and take a look at their behavior

```
CUSTOMER SEGMENT ANALYSIS ===
Detailed Customer Segment Analysis:
                     Count  Avg_Spent  Median_Spent  Total_Spent  Std_Spent  \
customer_segment
At Risk                925     517.08        410.15    478300.44     540.23
Champions             1703    6759.63       3365.22  11511655.36   22092.30
Lost Customers        1291     237.69        204.24    306861.37     159.13
Loyal Customers        931    1517.45       1190.23   1412747.39    1535.88
Potential Loyalists    969     893.80        717.21    866092.69    1156.20

                     Avg_Orders  Median_Orders  Std_Orders  Avg_Order_Value  \
customer_segment
At Risk                    1.78            2.0        0.83            33.03
Champions                 14.81           10.0       19.70            32.77
Lost Customers             1.14            1.0        0.35            24.08
Loyal Customers            4.65            4.0        2.73            29.62
Potential Loyalists        2.93            3.0        1.36            29.98

                     Median_Order_Value  …  Avg_Recency  Median_Recency  \
customer_segment                         …
At Risk                           16.59  …       257.02           240.0
Champions                         18.18  …        35.18            18.0
Lost Customers                    16.08  …       456.06           448.0
Loyal Customers                   16.89  …       109.49            64.0
Potential Loyalists               17.13  …       179.77           119.0

                     Std_Recency  Avg_Tenure_Months  Median_Tenure_Months  \
customer_segment
At Risk                   185.01              11.52                 12.65
Champions                  50.78              18.88                 21.29
Lost Customers            169.41              15.49                 15.60
Loyal Customers           120.37              15.52                 17.81
Potential Loyalists       161.99              13.51                 14.49

                     Avg_Purchase_Freq  Median_Purchase_Freq  \
customer_segment
At Risk                         184.07                 38.42
Champions                        13.12                  7.42
Lost Customers                  328.94                365.00
Loyal Customers                  13.18                  4.70
Potential Loyalists              54.52                  6.05
```

customer_segment	Avg_Unique_Products	Median_Unique_Products	Active_Rate
At Risk	31.65	25.0	0.30
Champions	171.87	130.0	0.91
Lost Customers	18.12	14.0	0.03
Loyal Customers	79.39	63.0	0.62
Potential Loyalists	49.12	39.0	0.45

[5 rows x 21 columns]

SEGMENT IMPACT ANALYSIS

Lost Customers:

- Customer Share: 1,291 customers (22.2%)
- Revenue Share: $306,861.37 (2.1%)
- Revenue per Customer: $237.69
- Active Rate: 2.7%
- Avg Recency: 456 days

Champions:

- Customer Share: 1,703 customers (29.3%)
- Revenue Share: $11,511,655.36 (79.0%)
- Revenue per Customer: $6759.63
- Active Rate: 91.0%
- Avg Recency: 35 days

Loyal Customers:

- Customer Share: 931 customers (16.0%)
- Revenue Share: $1,412,747.39 (9.7%)
- Revenue per Customer: $1517.45
- Active Rate: 61.9%
- Avg Recency: 109 days

At Risk:

- Customer Share: 925 customers (15.9%)
- Revenue Share: $478,300.44 (3.3%)
- Revenue per Customer: $517.08
- Active Rate: 29.8%
- Avg Recency: 257 days

Potential Loyalists:

- Customer Share: 969 customers (16.7%)
- Revenue Share: $866,092.69 (5.9%)
- Revenue per Customer: $893.80
- Active Rate: 44.6%
- Avg Recency: 180 days

4.0.1 Segment-Level Insights

1. Champions (29.3% of customers, 79% of revenue)

- They are the lifeblood of the business: small in number but extremely high spenders (~$6,760 each).
- Extremely active (91% active rate, recency ~35 days), with large and frequent orders (15 orders on average, 171 unique products).
- **Risk:** Heavy dependence on this group → losing even a small % will hit revenue hard.

2. Loyal Customers (16% of customers, 9.7% of revenue)

- Mid-value segment, still engaged (62% active, recency ~109 days).
- Spend per customer ~$1,517, avg. 5 orders, ~79 unique products.
- They are prime candidates to **upgrade into Champions** with personalized offers and upsell campaigns.

3. Potential Loyalists (16.7% of customers, 5.9% of revenue)

- Spend less (~$894 per customer) and engage moderately (45% active, recency ~180 days).
- Frequency is lower (3 orders, ~49 unique products).
- **Opportunity:** nurture with loyalty incentives, bundles, and cross-sell nudges. They're sitting in the middle, could move upward or churn.

4. At Risk (15.9% of customers, 3.3% of revenue)

- Weak engagement (30% active, recency ~257 days), low spend (~$517).
- Orders are infrequent (2 orders on average, ~32 unique products).
- They still have some recent touchpoints, so **reactivation campaigns** may save part of this group.

5. Lost Customers (22.2% of customers, only 2.1% of revenue)

- Lowest value segment (~$238 per customer), inactive for ~456 days.
- Very low activity (1 order, minimal product variety).
- **Impact is minimal on revenue**, so they're not worth aggressive win-back spend. Use low-cost automated reactivation if at all.

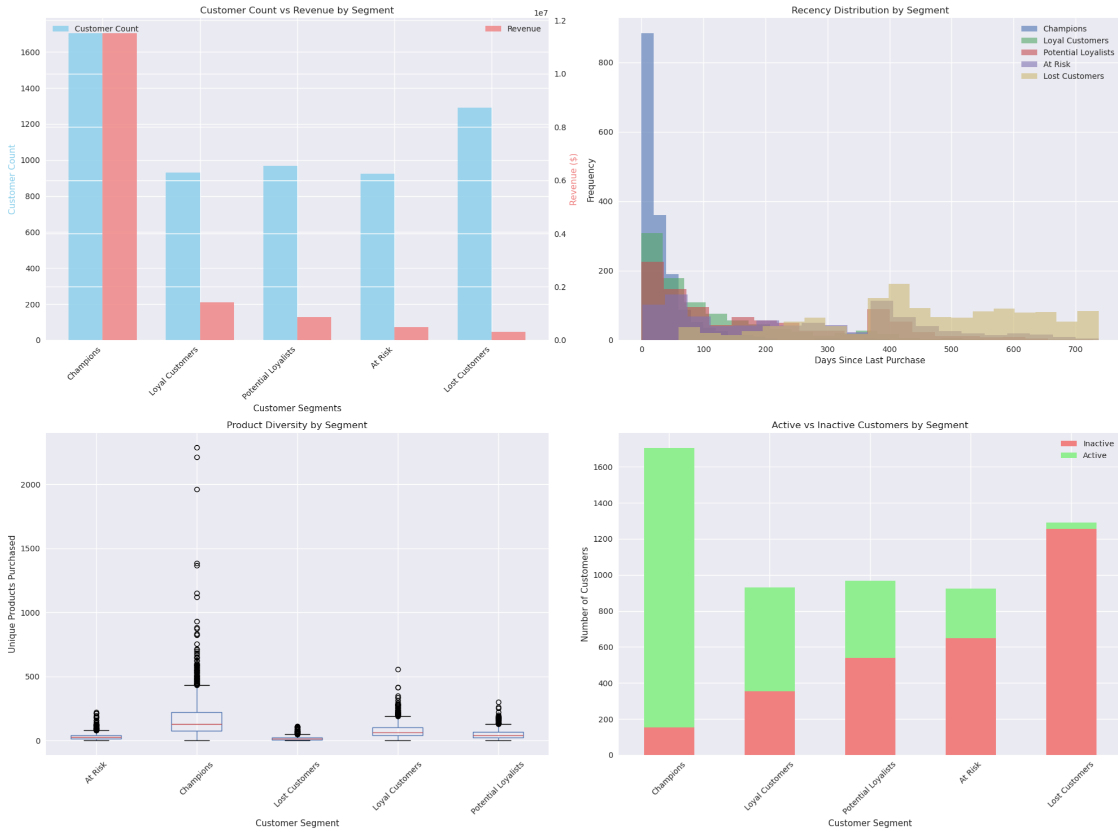
Takeaways

1. **Revenue is highly concentrated**: Champions + Loyal Customers = 45% of customers but 89% of revenue. Business depends heavily on these two groups.
2. **Retention gap is clear**: Potential Loyalists + At Risk represent 32% of the base. If they churn, they'll slide into "Lost" (already 22%). Retention efforts here could protect future revenue.
3. **Lost Customers are low value**: Losing them doesn't hurt much. Focus should remain on retaining and upgrading mid-tier and high-value customers.
4. **Cross-sell opportunity**: Champions buy many unique products (~172), while Potential Loyalists buy ~49. Bridging that gap through tailored product recommendations can accelerate their move upward.
5. **Recency gap**: Champions buy every ~1 month, Loyal every ~3 months, Potential Loyalists ~6 months. Tightening buying cycles through targeted offers could significantly increase

revenue.

Recommended Focus:

- **Protect & reward Champions** → VIP perks, early access, loyalty recognition.
- **Upgrade Loyal → Champions** → upselling high-value categories.
- **Nurture Potential Loyalists** → re-engagement campaigns, bundles, cross-sell into sticky products.
- **At-Risk Recovery** → targeted win-back with offering promo to make them stay.
- **Lost Customer** → don't spend much; focus resources upstream.



4.1 Segment Purchase Pattern & Product Preference

After we have know the characteristic of each segment, next we want to know if there is any product preference that distinct between these segment and if there is any purchase pattern from the quality customer that can be replicated to lower quality customer

```
PRODUCT CATEGORY ANALYSIS BY CUSTOMER SEGMENT
Transaction data with segments: 732,583 transactions
Available product categories: 9
Product categories: ['BEAUTY_PERSONAL', 'CHRISTMAS_HOLIDAY',
'FURNITURE_STORAGE', 'GARDEN_OUTDOOR', 'HOME_DECOR', 'KITCHEN_FOOD_UTENSIL',
```

```
'STATIONERY_OFFICE', 'TEXTILES_CLOTHING', 'TOYS_GAMES']

CATEGORY PREFERENCES BY SEGMENT

--- Champions Category Preferences ---
Top categories by revenue share:
```

- HOME_DECOR: 39.8% revenue, 11.9% penetration
- KITCHEN_FOOD_UTENSIL: 20.4% revenue, 11.8% penetration
- TEXTILES_CLOTHING: 9.4% revenue, 11.0% penetration
- BEAUTY_PERSONAL: 8.1% revenue, 11.7% penetration
- GARDEN_OUTDOOR: 6.1% revenue, 11.3% penetration

```
--- Loyal Customers Category Preferences ---
Top categories by revenue share:
```

- HOME_DECOR: 38.7% revenue, 12.9% penetration
- KITCHEN_FOOD_UTENSIL: 20.6% revenue, 12.6% penetration
- BEAUTY_PERSONAL: 8.8% revenue, 12.2% penetration
- TEXTILES_CLOTHING: 7.9% revenue, 10.8% penetration
- GARDEN_OUTDOOR: 5.8% revenue, 11.2% penetration

```
--- Potential Loyalists Category Preferences ---
Top categories by revenue share:
```

- HOME_DECOR: 37.5% revenue, 14.3% penetration
- KITCHEN_FOOD_UTENSIL: 21.9% revenue, 13.7% penetration
- BEAUTY_PERSONAL: 8.8% revenue, 12.8% penetration
- TEXTILES_CLOTHING: 7.7% revenue, 10.6% penetration
- GARDEN_OUTDOOR: 6.1% revenue, 11.3% penetration

```
--- At Risk Category Preferences ---
Top categories by revenue share:
```

- HOME_DECOR: 37.8% revenue, 16.3% penetration
- KITCHEN_FOOD_UTENSIL: 22.8% revenue, 15.0% penetration
- BEAUTY_PERSONAL: 8.6% revenue, 13.3% penetration
- GARDEN_OUTDOOR: 6.4% revenue, 11.1% penetration
- TEXTILES_CLOTHING: 6.2% revenue, 10.1% penetration

```
--- Lost Customers Category Preferences ---
Top categories by revenue share:
```

- HOME_DECOR: 38.8% revenue, 19.1% penetration
- KITCHEN_FOOD_UTENSIL: 21.0% revenue, 16.5% penetration
- BEAUTY_PERSONAL: 9.6% revenue, 13.7% penetration
- TEXTILES_CLOTHING: 6.9% revenue, 10.2% penetration
- GARDEN_OUTDOOR: 6.7% revenue, 11.1% penetration

What we learn: Category does not differentiate by loyalty tier; steady contribution but unlikely to drive transitions between segments.

1. **Home Decor = Core Engine**

   - Must always be protected in stock planning, pricing, and promotional visibility.
   - Best category to drive initial engagement and repeat purchases across all segments.

2. **Kitchen = Wide Appeal but High Churn Risk**
   - High penetration among At Risk/Lost customers suggests it can't retain on its own.
   - Bundle Kitchen products with Décor or Textiles to raise stickiness.

3. **Beauty = Entry Point**
   - Attracts both At Risk and Lost customers.
   - Cross-sell Beauty into higher-value categories (e.g., Décor, Kitchen) to prevent attrition.

4. **Textiles = Champion Driver**
   - Disproportionately stronger among Champions, indicating it is part of the upgrade path.
   - Push Textiles campaigns at Loyal and Potential Loyalists to encourage category expansion.

Call-to-Action:

- Anchor customer journeys on **Home Decor**, using it as the core of retention campaigns.
- Design **cross-sell flows**
- Focus on **Textiles growth** among mid-tier segments to create Champions.
- Use **low-cost add-ons** (Garden, Beauty) to increase basket size but don't expect them to drive loyalty shifts.

```
PRODUCT STICKINESS & RETENTION ANALYSIS
1. CATEGORY RETENTION CORRELATION:
Categories ranked by customer activity rate:
                      Avg_Recency  Avg_Orders  Avg_Spent  Avg_Lifespan  \
product_category
CHRISTMAS_HOLIDAY          132.82        8.11    3462.13        352.94
FURNITURE_STORAGE          151.12        8.43    3671.05        354.04
TOYS_GAMES                 159.30        7.83    3395.62        337.74
TEXTILES_CLOTHING          170.96        7.44    3154.70        324.50
STATIONERY_OFFICE          172.55        7.35    3096.01        322.51
GARDEN_OUTDOOR             175.22        7.29    3075.27        320.49
BEAUTY_PERSONAL            182.48        6.84    2854.14        304.99
KITCHEN_FOOD_UTENSIL       190.57        6.51    2692.57        290.50
HOME_DECOR                 195.44        6.26    2577.86        279.90

                      Active_Rate
product_category
CHRISTMAS_HOLIDAY            0.65
FURNITURE_STORAGE            0.60
TOYS_GAMES                   0.58
TEXTILES_CLOTHING            0.55
STATIONERY_OFFICE            0.55
GARDEN_OUTDOOR               0.54
BEAUTY_PERSONAL              0.53
KITCHEN_FOOD_UTENSIL         0.51
```

HOME_DECOR	0.50

2. EARLY PURCHASE CATEGORY IMPACT:
Impact of first purchase category on customer outcomes:

product_category	Customer_Count	Avg_Orders	Avg_CLV	Active_Rate
CHRISTMAS_HOLIDAY	277	5.66	2011.12	0.57
STATIONERY_OFFICE	379	5.02	1681.16	0.50
FURNITURE_STORAGE	161	5.63	2052.00	0.50
GARDEN_OUTDOOR	419	6.48	2154.29	0.50
HOME_DECOR	2247	6.14	2553.26	0.50
BEAUTY_PERSONAL	544	6.48	2670.06	0.49
TEXTILES_CLOTHING	471	6.51	2826.32	0.49
KITCHEN_FOOD_UTENSIL	1111	5.91	2325.82	0.48
TOYS_GAMES	210	6.96	4969.11	0.43

3. CHAMPIONS VS LOST CUSTOMERS CATEGORY PREFERENCES:
Categories favored by Champions vs Lost Customers:

Most Champions-favored categories:
- CHRISTMAS_HOLIDAY: Champions 10.6% vs Lost 5.4% (+5.2pp)
- FURNITURE_STORAGE: Champions 10.0% vs Lost 5.6% (+4.4pp)
- TOYS_GAMES: Champions 10.4% vs Lost 8.0% (+2.4pp)
- STATIONERY_OFFICE: Champions 11.3% vs Lost 10.4% (+0.8pp)
- TEXTILES_CLOTHING: Champions 11.0% vs Lost 10.2% (+0.8pp)

Most Lost-Customer-favored categories:
- BEAUTY_PERSONAL: Champions 11.7% vs Lost 13.7% (-2.0pp)
- KITCHEN_FOOD_UTENSIL: Champions 11.8% vs Lost 16.5% (-4.7pp)
- HOME_DECOR: Champions 11.9% vs Lost 19.1% (-7.2pp)

4. CATEGORY CROSS-SELLING PATTERNS:
Impact of category diversity on customer value:

categories_purchased	Customer_Count	Avg_CLV	Avg_Orders	Active_Rate
1	182	522.30	1.71	0.21
2	189	457.23	1.74	0.25
3	320	509.38	1.88	0.29
4	362	556.81	2.10	0.30
5	495	881.15	2.48	0.28
6	609	951.73	2.89	0.36
7	765	1176.24	3.85	0.40
8	999	1906.84	5.55	0.49
9	1898	5378.87	11.60	0.75

Optimal category diversity for CLV: 9 categories
CLV at optimal diversity: $5378.87

```
5. SEQUENTIAL PURCHASE PATTERNS:
Same-category repeat purchase rate: 40.7%

Top category transitions (First → Second purchase):
  • HOME_DECOR → HOME_DECOR: 196 customers
  • KITCHEN_FOOD_UTENSIL → KITCHEN_FOOD_UTENSIL: 81 customers
  • HOME_DECOR → KITCHEN_FOOD_UTENSIL: 62 customers
  • KITCHEN_FOOD_UTENSIL → HOME_DECOR: 57 customers
  • TEXTILES_CLOTHING → TEXTILES_CLOTHING: 41 customers
  • HOME_DECOR → TEXTILES_CLOTHING: 32 customers
  • HOME_DECOR → BEAUTY_PERSONAL: 28 customers
  • BEAUTY_PERSONAL → HOME_DECOR: 27 customers
  • HOME_DECOR → GARDEN_OUTDOOR: 24 customers
  • GARDEN_OUTDOOR → HOME_DECOR: 22 customers
```

4.1.1 1. Category Retention & Stickiness

- **High-Activity Categories** (best for retention):
  - **Christmas/Holiday, Furniture/Storage, Toys/Games** → Customers who buy here show higher order counts (~8+) and longer lifespans (~350 days).
  - These categories **correlate with "Champions"** – they bring stronger ongoing activity.
- **Weak Stickiness Categories**:
  - **Home Decor & Kitchen/Food** dominate revenue, but customers here are **less active** (avg. 6 orders, lifespan ~280 days).
  - They are big **acquisition funnels** but underperform in converting to loyal customers.

4.1.2 2. First Purchase Impact

- **High-Value First Purchases**:
  - Customers starting in **Toys/Games** have the **highest CLV** ($4,969) but poor retention (active rate 0.43).
  - **Kitchen/Food & Beauty/Personal** → balance between volume and stickiness; good entry points for loyalty.
- **Lower Value Onboarding**:
  - **Stationery & Garden** yield lower CLV (<$2,200) despite decent activity.

4.1.3 3. Champions vs Lost Customers

- **Champions buys** seasonal/niche: Christmas, Furniture, Toys.
- **Lost Customers skew** toward everyday staples: Home Décor, Kitchen, Beauty.

4.1.4 4. Cross-Selling & Category Diversity

- Strong positive relationship:
  - CLV grows dramatically with category diversity.

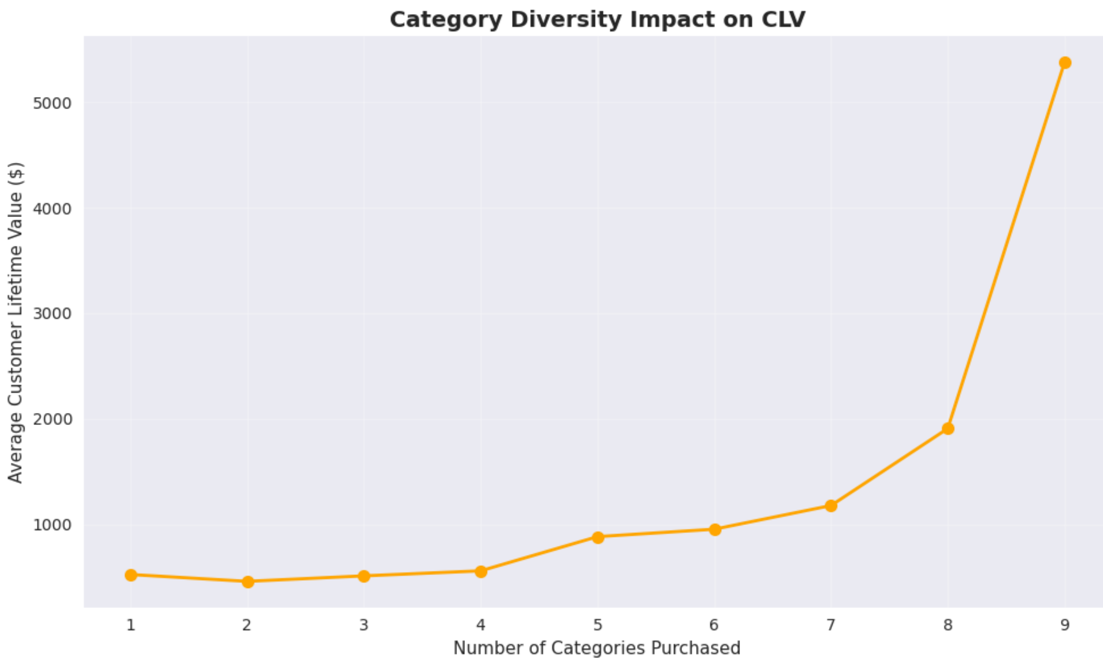
- Customers buying from **9 categories** are worth ~**$5,379 CLV**, 10× higher than single-category buyers.

- Active rate rises from **21% (1 category) → 75% (9 categories)**.

Cross-category engagement is the single strongest lever for retention and value. Campaigns must actively push multi-category shopping.

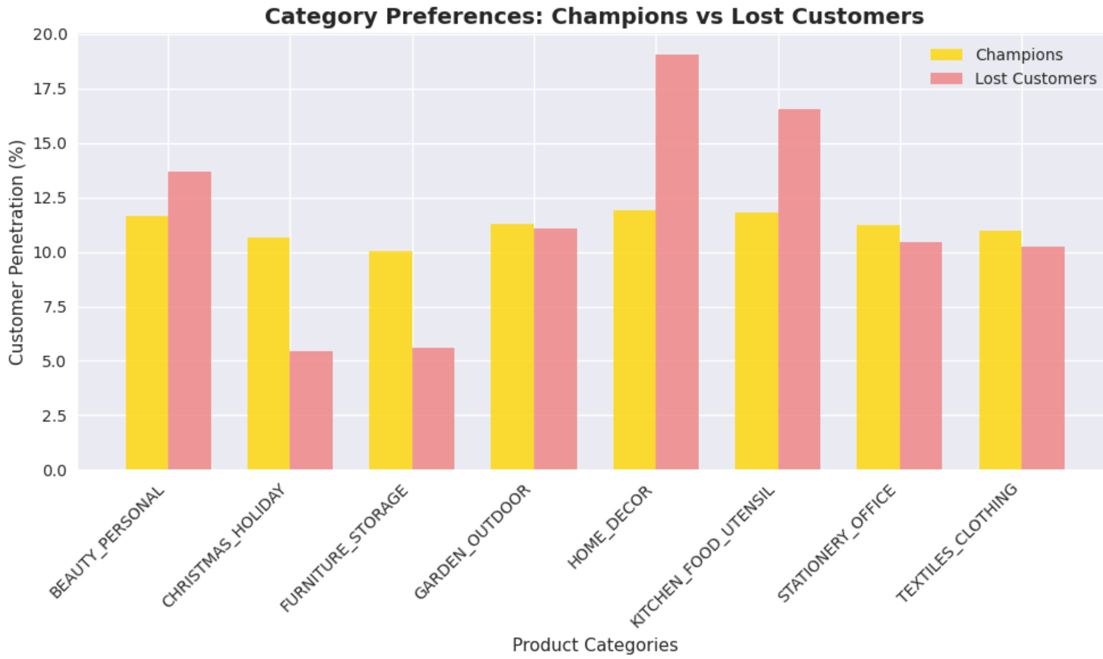
4.1.5 5. Sequential Purchase Patterns

- **Same-category repeat rate = 41%** → customers like to re-buy within the same category (esp. Home Decor, Kitchen).



Here we can see from the chart that Customer Lifetime Value (CLV) grows almost exponentially as customers purchase across more categories. For single-category buyers, CLV is very low (~$500), and it stays relatively flat up to 4 categories. But starting from 5+ categories, CLV rises steeply, reaching nearly $5,400 when customers buy across all 9 categories.

Customers who diversify their purchases are not just buying more often, but also staying active longer, which compounds their lifetime value.



Everyday categories (Home Decor, Kitchen, Beauty) are essential for acquisition but insufficient for retention. Without cross-sell into higher-stickiness categories, these customers churn.

Seasonal/niche categories (Christmas, Furniture, Toys) are associated with Champions and should be used as "retention hooks."

5 Phase 3: Machine Learning for Churn Prediction

In this section, we'll develop and evaluate machine learning models to predict customer churn

1. **Data Preparation & Feature Engineering** - Create predictive features without data leakage
2. **Data Splitting Strategy** - Split the data, predict the last 90 days of data
3. **Model Training** - Train multiple algorithms (Logistic Regression, Random Forest, XGBoost)
4. **Model Evaluation & Comparison** - Compare performance and select best model
5. **Final Predictions** - Generate churn scores and recommendations

5.1 Data Preparation & Feature Engineering

5.1.1 Import Required Libraries

5.1.2 Temporal Data Split (Preventing Data Leakage)

```
Data Timeline:
  • Full period: 2009-12-01 to 2011-12-09 (738 days)
  • Training period: 2009-12-01 to 2011-07-14
  • Validation period: 2011-07-14 to 2011-12-09
  • Prediction window: 90 days
```

- Active threshold: 90 days (exclude already churned customers)

Training Data (Excluding Already Churned):
- All training transactions: 531,153
- All customers in training period: 5,022
- Active customers at split date: 1,923
- Already churned (excluded): 3,099
- Active training transactions: 347,854

Churn Target Definition (Active Customers Only):
- Active customers at split: 1,923
- Customers active after split: 1,436
- Customers who churned after split: 487
- Churn rate (among active): 25.3%

5.1.3 Feature Engineering

Create predictive features using only data available up to the split date.

Feature Engineering Complete:
- Total customers: 1,923
- Total features: 24
- Churn rate: 25.3%

Feature Categories:
- Volume features: total_orders, total_transactions, total_quantity
- Monetary features: total_spent, avg_order_value, monetary_per_day
- Behavioral features: purchase_frequency, product_diversity, category_diversity
- Temporal features: days_since_first, days_since_last, customer_lifespan
- Early behavior: early_revenue, early_orders, early_products, early_categories

	customer_id	total_orders	first_purchase	last_purchase_in_training	\
0	12347.0	5	2010-10-31 14:20:00	2011-06-09 13:01:00	
1	12353.0	2	2010-10-27 12:44:00	2011-05-19 17:47:00	
2	12354.0	1	2011-04-21 13:11:00	2011-04-21 13:11:00	
3	12355.0	2	2010-05-21 11:59:00	2011-05-09 13:49:00	
4	12358.0	4	2009-12-08 07:59:00	2011-07-12 10:04:00	

	total_transactions	total_spent	avg_order_value	spend_volatility	\
0	142	2817.48	19.841408	21.527576	
1	24	406.76	16.948333	8.799589	
2	58	1079.40	18.610345	8.679742	
3	35	947.61	27.074571	22.110475	
4	68	2923.87	42.998088	64.174167	

```
   total_quantity  avg_quantity_per_order  …  monetary_per_day  \
0            1822               12.830986  …         11.005781
1             212                8.833333  …          1.564462
2             530                9.137931  …         12.850000
3             543               15.514286  …          2.261599
4             924               13.588235  …          5.015214

   product_diversity  category_diversity  avg_days_between_orders  \
0          16.833333                 0.9                36.666667
1           7.666667                 0.5                68.000000
2          29.000000                 0.8                 0.000000
3          11.666667                 0.7               117.666667
4          11.000000                 0.6               116.200000

   early_revenue  early_orders  early_products  early_categories  \
0         611.53             1              40                 7
1         317.76             1              20                 5
2        1079.40             1              58                 8
3         488.21             1              22                 6
4        1429.83             1              17                 5

   early_avg_order_value  churned
0              15.288250        0
1              15.888000        1
2              18.610345        1
3              22.191364        1
4              84.107647        0

[5 rows x 26 columns]
```

5.2 Train Test Data Splitting Strategy

Dataset Prepared:

- Feature matrix shape: (1923, 22)
- Target vector shape: (1923,)
- Features: 22
- Positive class (churned): 487 (25.3%)
- Negative class (active): 1,436 (74.7%)

Train-Validation Split:

- Training set: 1,538 customers (80.0%)
- Validation set: 385 customers (20.0%)
- Training churn rate: 25.3%
- Validation churn rate: 25.5%

Data ready for model training

- X_train: (1538, 22) (original)
- X_train_scaled: (1538, 22) (standardized)

- `X_val: (385, 22) (original)`
- `X_val_scaled: (385, 22) (standardized)`

5.3 Model Training

Train and tune three different machine learning algorithms to predict customer churn.

5.3.1 Logistic Regression

```
Logistic Regression Results:
```
- `AUC-ROC: 0.786`
- `Average Precision: 0.549`
- `F1-Score: 0.542`
- `CV AUC: 0.769 ± 0.029`

```
Top 5 Most Important Features (Logistic Regression):
   6. total_quantity            Decreases churn risk (coef: -1.515)
   2. total_transactions        Decreases churn risk (coef: -1.144)
   3. total_spent               Decreases churn risk (coef: -0.995)
   17. avg_days_between_orders   Increases churn risk (coef: +0.677)
   8. unique_products           Increases churn risk (coef: +0.618)

Logistic Regression trained successfully
```

Logistic Regression was chosen first because it's **simple, interpretable, and fast**. It sets a clear baseline before moving to more complex models (like Random Forest or XGBoost) which may improve accuracy but sacrifice interpretability.

1. **Class Imbalance**
   - Churn datasets are often imbalanced (fewer churners than non-churners).
   - We used `class_weight='balanced'` so the model pays equal attention to both classes, avoiding bias toward the majority.
2. **Feature Scaling**
   - Logistic Regression is sensitive to feature scale.
   - Input features were standardized (`X_train_scaled`) to ensure coefficients are comparable and model converges efficiently.
3. **Interpretability**
   - One strength of Logistic Regression is **transparent coefficients**.
   - Positive coefficients → increase churn risk; negative → reduce churn risk. This helps translate results into business insights (e.g., more orders reduce churn).
4. **Evaluation Metrics**
   - Used multiple metrics to balance perspectives:
     - **AUC-ROC (0.786):** ability to rank churn vs non-churn.
     - **Average Precision (0.549):** robustness in imbalanced setting.
     - **F1-score (0.542):** balance of precision and recall.

- Cross-validation confirmed stability (AUC ~0.769 ± 0.029).

5.3.2 Random Forest

```
Random Forest Results:
  • AUC-ROC: 0.765
  • Average Precision: 0.514
  • F1-Score: 0.525
  • CV AUC: 0.772 ± 0.030

Top 5 Most Important Features (Random Forest):
   3. total_spent               (importance: 0.136)
   6. total_quantity            (importance: 0.117)
   1. total_orders              (importance: 0.083)
   14. monetary_per_day          (importance: 0.081)
   2. total_transactions        (importance: 0.076)

Random Forest trained successfully
```

Unlike Logistic Regression, Random Forest handles raw feature scales directly. More flexible in capturing **nonlinear patterns** and feature interactions.

1. **Class Imbalance**
   - `class_weight='balanced'` was applied to prevent bias toward the majority (non-churners).
2. **Hyperparameters**
   - `max_depth=10, min_samples_split=20, min_samples_leaf=10` → constraints added to **avoid overfitting** while keeping interpretability of feature importance.
   - `n_estimators=100` ensures stability of predictions.
3. **Performance**
   - **AUC-ROC = 0.765**, **AP = 0.514**, **F1 = 0.525**.
   - Slightly weaker than Logistic Regression (AUC 0.786), but **cross-validation AUC ~0.772 ± 0.030** shows stable generalization.

5.3.3 XGBoost

```
XGBoost Results:
  • AUC-ROC: 0.739
  • Average Precision: 0.465
  • F1-Score: 0.483
  • CV AUC: 0.739 ± 0.019

Top 5 Most Important Features (XGBoost):
   3. total_spent               (importance: 0.114)
   6. total_quantity            (importance: 0.080)
   17. avg_days_between_orders   (importance: 0.060)
   1. total_orders              (importance: 0.057)
```

```
   13. purchase_frequency        (importance: 0.049)
```

```
XGBoost trained successfully
```

XGBoost builds trees sequentially, correcting mistakes from previous trees. This allows it to capture nonlinear patterns and interactions much better than Logistic Regression or Random Forest.

1. **Class Imbalance**
   - Used scale_pos_weight = (# non-churners / # churners) to balance churn prediction, ensuring the model does not just predict the majority class.
2. **Regularization & Stability**
   - Parameters (max_depth=6, subsample=0.8, colsample_bytree=0.8) help reduce overfitting.
   - learning_rate=0.1 controls how fast the model adapts, trading off speed vs generalization.

Overall the predictive power is slightly weaker compared to Logistic Regression (AUC ~0.786) and roughly on par with Random Forest. A caveat is that XGBoost is more data-hungry and parameter-sensitive. With richer behavioral or marketing features (campaign response, engagement signals), it may show stronger advantages. For now, Logistic Regression remains the best balance of accuracy and interpretability.

5.4 Model Evaluation and Comparison

```
                 Model    Recall  Precision  F1-Score   AUC-ROC  \
0  Logistic Regression  0.755102   0.422857  0.542125  0.785999
1        Random Forest  0.653061   0.438356  0.524590  0.765022
2              XGBoost  0.500000   0.466667  0.482759  0.739245

   Average Precision
0           0.548980
1           0.513514
2           0.465052


BEST MODEL: Logistic Regression
   • Best Recall: 0.755 (75.5% of churners detected)
   • AUC-ROC: 0.786
   • Model Type: Linear
   • Selection Criteria: Highest recall for churn detection

DETAILED EVALUATION - Logistic Regression:

Classification Report:
              precision    recall  f1-score   support

      Active       0.89      0.65      0.75       287
     Churned       0.42      0.76      0.54        98
```

```
    accuracy                           0.68       385
   macro avg       0.65      0.70      0.65       385
weighted avg       0.77      0.68      0.70       385


Confusion Matrix:
                 Predicted
Actual           Active  Churned
Active             186     101
Churned             24      74
```

Business Metrics:

- RECALL (Sensitivity): 75.5% of churners identified
- Precision: 42.3% of churn predictions are correct
- Specificity: 64.8% of active customers correctly identified

PREDICTING CHURN FOR NEXT 3 MONTHS

```
Analysis date: 2011-12-09
Predicting churn for 3-month period: 2011-12-09 to 2012-03-09

Creating features for prediction (matching training features)…
Calculating early behavior features…
Features created for prediction: ['total_orders', 'total_transactions',
'total_spent', 'avg_order_value', 'spend_volatility', 'total_quantity',
'avg_quantity_per_order', 'unique_products', 'unique_categories',
'days_since_first', 'days_since_last', 'customer_lifespan',
'purchase_frequency', 'monetary_per_day', 'product_diversity',
'category_diversity', 'avg_days_between_orders', 'early_revenue',
'early_orders', 'early_products', 'early_categories', 'early_avg_order_value']
Shape of prediction dataset: (5819, 22)
Excluding 'Lost Customers' from churn prediction list…
Customers before excluding 'Lost Customers': 5,819
Customers after excluding 'Lost Customers': 4,528

CHURN PREDICTION SUMMARY (Next 3 Months) - HIGH RISK ONLY (>80%):
Total customers analyzed (excluding Lost Customers): 4,528
High risk customers (>80%): 1,556
High risk rate: 34.4%
```

Risk Distribution (High Risk Only):

- High Risk: 527 customers (33.9%)
- Critical Risk: 1,029 customers (66.1%)

```
Segment-wise High Risk Distribution:
                     Avg_Churn_Prob  Customer_Count
customer_segment
```

```
At Risk                     0.942          690
Potential Loyalists         0.918          599
Loyal Customers             0.884          254
Champions                   0.878           13

High-Value Customers at High Risk (Top 20):
      customer_id     customer_segment  churn_probability  total_spent  \
4412      18139.0            Champions           0.999134      8438.34
2727      16000.0            Champions           0.999069     12393.70
1944      14938.0              At Risk           0.998777      1757.31
3269      16716.0              At Risk           0.998428      1248.48
3744      17305.0              At Risk           0.997938      2135.46
2605      15823.0  Potential Loyalists           0.997655      3217.21
315       12742.0              At Risk           0.996891      1185.02
3539      17039.0              At Risk           0.996403      1954.99
1864      14831.0  Potential Loyalists           0.996378      1440.12
2538      15736.0              At Risk           0.995937      1682.17
2813      16118.0              At Risk           0.995643      3997.73
260       12671.0              At Risk           0.994823      2622.48
650       13205.0              At Risk           0.994502      2803.20
2295      15413.0      Loyal Customers           0.994470      6798.72
1956      14956.0              At Risk           0.994204      1325.00
910       13543.0              At Risk           0.994127      1439.61
4348      18051.0  Potential Loyalists           0.994015      1863.48
3864      17448.0      Loyal Customers           0.993846     13928.02
437       12911.0              At Risk           0.993644      1651.72
747       13337.0              At Risk           0.993296      1550.06

      total_orders  days_since_last  risk_category
4412             6               17  Critical Risk
2727             3                2  Critical Risk
1944             2              561  Critical Risk
3269             1              630  Critical Risk
3744             1              648  Critical Risk
2605             2              728  Critical Risk
315              1              625  Critical Risk
3539             1              603  Critical Risk
1864             3              679  Critical Risk
2538             2              434  Critical Risk
2813             1              652  Critical Risk
260              1              605  Critical Risk
650              1              442  Critical Risk
2295             5              691  Critical Risk
1956             1              420  Critical Risk
910              2              633  Critical Risk
4348             7              633  Critical Risk
3864            41              528  Critical Risk
```

```
437             1            551  Critical Risk
747             1            545  Critical Risk


Total revenue at risk from high-value customers: $73,432.82

High Risk Customers Summary:
Count: 1,556 customers
Total spent at risk: $1,522,389.17
Average spend per at-risk customer: $978.40

High risk churn prediction complete. Results saved for 1,556 customers.
```

Here we can see from the model comparison that Logistic Regression gave the best balance of accuracy (AUC ~0.786) and interpretability, while Random Forest (AUC ~0.765) and XGBoost (AUC ~0.739) provided additional validation but did not significantly outperform. All three models highlight the same key churn drivers: higher spend, quantity, and transaction frequency reduce churn, while longer purchase gaps increase churn risk.

The insight here is that churn in this business is strongly linked to engagement depth and consistency. Customers who are active across multiple purchases and categories remain sticky, while those who disengage for long periods are at high risk of churn.

What we learn is that our models are reliable enough to flag customers at risk, but precision is not perfect. This creates an opportunity to over-predict churn deliberately: treating a broader group of customers as "at risk" may waste some effort on false positives, but it ensures we don't miss truly at-risk customers.

The next step is to use churn scores to over-identify at-risk customers, then reactivate them with sticky-category recommendations.

6 Phase 4: Recommendation System

In this phase, we design a simple recommendation system aligned with three key business scenarios, ensuring recommendations support revenue growth, customer stickiness, and churn reactivation.

We target different segment with different scenario

```
* Scenario 1: Champions and Loyal Customer (Drive Revenue)
* Scenario 2: At Risk and Potential Loyal (Stickiness)
* Scenario 3: Churn Prevention
```

Scenario 1: Drive Revenue

- **Logic:**
    1. Identify customer's **favorite category** (highest spending).
    2. Retrieve **top-selling products** in that category.
    3. Recommend **3 products the customer has not purchased**.
- **Goal:** Increase revenue by deepening spend in customer's preferred area.

Scenario 2: Cross-Selling (Stickiness)

- **Logic:**
  1. Start from customer's **favorite category**.
  2. Identify **related categories** (via correlation analysis).
  3. Recommend **1 top product from each related category** (3 in total).
- **Goal:** Broaden category engagement, driving **higher stickiness and CLV**.

Scenario 3: Churn Prevention

- **Logic:**
  1. Use churn model to identify **high-risk customers**.
  2. Focus on **high-stickiness categories** (>60% retention).
  3. Recommend **Top 3 products they haven't tried** from these categories.
  4. Offer promo to win back high risk customers
- **Goal:** Win back disengaged customers with proven retention drivers.

And we will also consider seasonal item * Exclude **Christmas/Holiday** products for recommendations outside the festive period. * Focus on **year-round categories** for sustainable engagement.

6.0.1 Data Preparation for Recommendation System

First, let's analyze category correlations and prepare the data needed for our recommendation scenarios.

```
=== RECOMMENDATION SYSTEM DATA PREPARATION ===

1. CATEGORY CORRELATION ANALYSIS:
Category correlation matrix:
```

```
product_category      BEAUTY_PERSONAL  CHRISTMAS_HOLIDAY  FURNITURE_STORAGE  \
product_category
BEAUTY_PERSONAL                 1.000              0.668              0.803
CHRISTMAS_HOLIDAY               0.668              1.000              0.528
FURNITURE_STORAGE               0.803              0.528              1.000
GARDEN_OUTDOOR                  0.806              0.425              0.693
HOME_DECOR                      0.874              0.501              0.791
KITCHEN_FOOD_UTENSIL            0.875              0.704              0.795
STATIONERY_OFFICE               0.779              0.598              0.724
TEXTILES_CLOTHING               0.633              0.595              0.514
TOYS_GAMES                      0.657              0.667              0.551

product_category      GARDEN_OUTDOOR  HOME_DECOR  KITCHEN_FOOD_UTENSIL  \
product_category
BEAUTY_PERSONAL                0.806       0.874                 0.875
CHRISTMAS_HOLIDAY              0.425       0.501                 0.704
FURNITURE_STORAGE              0.693       0.791                 0.795
GARDEN_OUTDOOR                 1.000       0.960                 0.723
HOME_DECOR                     0.960       1.000                 0.820
KITCHEN_FOOD_UTENSIL           0.723       0.820                 1.000
```

```
STATIONERY_OFFICE              0.675       0.761                  0.806
TEXTILES_CLOTHING              0.446       0.544                  0.708
TOYS_GAMES                     0.467       0.565                  0.757

product_category      STATIONERY_OFFICE  TEXTILES_CLOTHING  TOYS_GAMES
product_category
BEAUTY_PERSONAL                   0.779              0.633       0.657
CHRISTMAS_HOLIDAY                 0.598              0.595       0.667
FURNITURE_STORAGE                 0.724              0.514       0.551
GARDEN_OUTDOOR                    0.675              0.446       0.467
HOME_DECOR                        0.761              0.544       0.565
KITCHEN_FOOD_UTENSIL              0.806              0.708       0.757
STATIONERY_OFFICE                 1.000              0.605       0.612
TEXTILES_CLOTHING                 0.605              1.000       0.762
TOYS_GAMES                        0.612              0.762       1.000

2. CATEGORY STICKINESS ANALYSIS:
Category stickiness (retention rate):

               category  retention_rate
0     CHRISTMAS_HOLIDAY        0.654719
5     FURNITURE_STORAGE        0.595630
7            TOYS_GAMES        0.576903
8     TEXTILES_CLOTHING        0.554923
6     STATIONERY_OFFICE        0.546602
4        GARDEN_OUTDOOR        0.540821
1       BEAUTY_PERSONAL        0.529591
3  KITCHEN_FOOD_UTENSIL        0.509516
2            HOME_DECOR        0.502674

High-stickiness categories (>60% retention): ['CHRISTMAS_HOLIDAY']

3. PRODUCT PERFORMANCE ANALYSIS:
Product performance calculated for 4755 products
Top 5 products by popularity score:

     product_id      product_category  total_revenue  unique_customers  \
4215     85123A            HOME_DECOR      148184.57              1395
1652      22423  KITCHEN_FOOD_UTENSIL      125625.55              1121
3938      84879            HOME_DECOR      121829.55               995
4193     85099B     TEXTILES_CLOTHING      119378.46               930
3163      47566  KITCHEN_FOOD_UTENSIL       95829.48               874

      popularity_score
4215          1.000000
1652          0.830092
3938          0.778593
4193          0.750031
```

```
3163          0.638623
```

```
4. CUSTOMER PURCHASE HISTORY:
Purchase history compiled for 5819 customers

5. SEASONAL PRODUCT IDENTIFICATION:
Seasonal categories to filter: ['CHRISTMAS_HOLIDAY']
Current month: 12, Is festive period: True
```

6.0.2 Scenario 1: Drive Revenue (Champions and Loyal Customers)

Logic: 1. Identify customer's favorite category (highest spending) 2. Retrieve top-selling products in that category 3. Recommend 3 products the customer has not purchased

Goal: Increase revenue by deepening spend in customer's preferred area.

```
=== SCENARIO 1: DRIVE REVENUE RECOMMENDATIONS ===
Target: Champions and Loyal Customers
Generating recommendations for 2634 Champions and Loyal Customers…

Scenario 1 Results:
```

- Customers targeted: 2,634
- Recommendations generated: 7,902
- Average recommendations per customer: 3.0
- Categories recommended: 9

```
Top categories in Scenario 1 recommendations:
```

- HOME_DECOR: 6231 recommendations
- KITCHEN_FOOD_UTENSIL: 987 recommendations
- TEXTILES_CLOTHING: 345 recommendations
- CHRISTMAS_HOLIDAY: 117 recommendations
- STATIONERY_OFFICE: 66 recommendations

```
Sample Scenario 1 Recommendations:
```

	customer_id	customer_segment	recommended_product_id	recommended_category \
0	12347.0	Champions	85123A	HOME_DECOR
1	12347.0	Champions	84879	HOME_DECOR
2	12347.0	Champions	22469	HOME_DECOR
3	12348.0	Loyal Customers	22423	KITCHEN_FOOD_UTENSIL
4	12348.0	Loyal Customers	47566	KITCHEN_FOOD_UTENSIL
5	12348.0	Loyal Customers	21212	KITCHEN_FOOD_UTENSIL
6	12349.0	Champions	85123A	HOME_DECOR
7	12349.0	Champions	22469	HOME_DECOR
8	12349.0	Champions	22138	HOME_DECOR
9	12352.0	Champions	85123A	HOME_DECOR

	reason
0	Top product in favorite category (HOME_DECOR)

```
1      Top product in favorite category (HOME_DECOR)
2      Top product in favorite category (HOME_DECOR)
3  Top product in favorite category (KITCHEN_FOOD…
4  Top product in favorite category (KITCHEN_FOOD…
5  Top product in favorite category (KITCHEN_FOOD…
6      Top product in favorite category (HOME_DECOR)
7      Top product in favorite category (HOME_DECOR)
8      Top product in favorite category (HOME_DECOR)
9      Top product in favorite category (HOME_DECOR)

Scenario 1 implementation completed!
```

6.0.3 Scenario 2: Cross-Selling for Stickiness (At Risk and Potential Loyalists)

Logic: 1. Start from customer's favorite category 2. Identify related categories (via correlation analysis) 3. Recommend 1 top product from each related category (3 in total)

Goal: Broaden category engagement, driving higher stickiness and CLV.

```
=== SCENARIO 2: CROSS-SELLING FOR STICKINESS RECOMMENDATIONS ===
Target: At Risk and Potential Loyalists
Generating recommendations for 1894 At Risk and Potential Loyalists…

Scenario 2 Results:
  • Customers targeted: 1,894
  • Recommendations generated: 2,938
  • Average recommendations per customer: 1.6
  • Categories recommended: 9

Top categories in Scenario 2 recommendations:
  • FURNITURE_STORAGE: 821 recommendations
  • TOYS_GAMES: 594 recommendations
  • STATIONERY_OFFICE: 465 recommendations
  • GARDEN_OUTDOOR: 393 recommendations
  • BEAUTY_PERSONAL: 326 recommendations

Recommendation reasons breakdown:
  • New category recommendations: 2903
  • High retention category recommendations: 35

Sample Scenario 2 Recommendations:

   customer_id customer_segment recommended_product_id recommended_category  \
0      12353.0          At Risk                  21754    FURNITURE_STORAGE
1      12353.0          At Risk                  21791           TOYS_GAMES
2      12353.0          At Risk                  23298       GARDEN_OUTDOOR
3      12355.0          At Risk                  21754    FURNITURE_STORAGE
4      12361.0          At Risk                  21755      BEAUTY_PERSONAL
5      12361.0          At Risk                  48138    STATIONERY_OFFICE
```

```
6      12361.0          At Risk                22086    CHRISTMAS_HOLIDAY
7      12363.0          At Risk                21755      BEAUTY_PERSONAL
8      12363.0          At Risk                21791           TOYS_GAMES
9      12363.0          At Risk               85099B    TEXTILES_CLOTHING

                                            reason
0  Cross-sell from KITCHEN_FOOD_UTENSIL to FURNIT…
1  Cross-sell from KITCHEN_FOOD_UTENSIL to TOYS_G…
2  Cross-sell from KITCHEN_FOOD_UTENSIL to GARDEN…
3  Cross-sell from KITCHEN_FOOD_UTENSIL to FURNIT…
4  Cross-sell from TEXTILES_CLOTHING to BEAUTY_PE…
5  Cross-sell from TEXTILES_CLOTHING to STATIONER…
6  Cross-sell from TEXTILES_CLOTHING to CHRISTMAS…
7  Cross-sell from CHRISTMAS_HOLIDAY to BEAUTY_PE…
8  Cross-sell from CHRISTMAS_HOLIDAY to TOYS_GAME…
9  Cross-sell from CHRISTMAS_HOLIDAY to TEXTILES_…


Scenario 2 implementation completed!
```

6.0.4 Scenario 3: Churn Prevention (High-Risk Customers)

Logic: 1. Use churn model to identify high-risk customers 2. Focus on high-stickiness categories (>60% retention) 3. Recommend top 3 products they haven't tried from these categories 4. Offer promotional incentives to win back high risk customers

Goal: Win back disengaged customers with proven retention drivers.

```
=== SCENARIO 3: CHURN PREVENTION RECOMMENDATIONS ===
Target: Customers Predicted to Churn (90-day prediction)

1. USING EXISTING CHURN PREDICTIONS:
   • Using customers from next_3_month_churn_list
   • Customers predicted to churn (90-day window): 1556
   • Average churn probability: 0.923

2. GENERATING CHURN PREVENTION RECOMMENDATIONS:
Targeting 1556 customers predicted to churn (90-day window)

Scenario 3 Results:
   • Customers targeted: 1,556
   • Recommendations generated: 4,668
   • Average recommendations per customer: 3.0
   • Categories recommended: 5

Top categories in Scenario 3 recommendations:
   • FURNITURE_STORAGE: 1497 recommendations
   • TOYS_GAMES: 1446 recommendations
   • TEXTILES_CLOTHING: 1277 recommendations
```

- STATIONERY_OFFICE: 265 recommendations
- GARDEN_OUTDOOR: 183 recommendations

Sample Scenario 3 Recommendations:

	customer_id	recommended_product_id	recommended_category	churn_probability
0	17945.0	21754	FURNITURE_STORAGE	0.999939
1	17945.0	21791	TOYS_GAMES	0.999939
2	17945.0	85099B	TEXTILES_CLOTHING	0.999939
3	15959.0	21754	FURNITURE_STORAGE	0.999933
4	15959.0	21791	TOYS_GAMES	0.999933
5	15959.0	85099B	TEXTILES_CLOTHING	0.999933
6	15794.0	21791	TOYS_GAMES	0.999817
7	15794.0	23298	GARDEN_OUTDOOR	0.999817
8	15794.0	21754	FURNITURE_STORAGE	0.999817
9	13526.0	21754	FURNITURE_STORAGE	0.999781

	risk_level	promotional_offer
0	High	25% discount + free shipping
1	High	25% discount + free shipping
2	High	25% discount + free shipping
3	High	25% discount + free shipping
4	High	25% discount + free shipping
5	High	25% discount + free shipping
6	High	25% discount + free shipping
7	High	25% discount + free shipping
8	High	25% discount + free shipping
9	High	25% discount + free shipping

Scenario 3 implementation completed!

6.0.5 Final Customer 360 Recommendation Table

Now let's create the final comprehensive table that combines all scenarios and shows each customer with their segment and top 3 product recommendations based on their appropriate scenario.

=== FINAL CUSTOMER 360 RECOMMENDATION TABLE ===

1. COMBINING ALL RECOMMENDATION SCENARIOS:
   - Scenario 1 (Revenue): 7902 recommendations
   - Scenario 2 (Stickiness): 2938 recommendations
   - Scenario 3 (Churn Prevention): 4668 recommendations
   - Total recommendations: 15508

1.5. CREATING PRODUCT LOOKUP TABLE:
   - Product lookup table created with 4604 products

2. CREATING CUSTOMER-LEVEL SUMMARY:

3. RECOMMENDATION SUMMARY STATISTICS:
Customers by primary scenario (hierarchy applied):

- Revenue Growth: 2367 customers
- Churn Prevention: 1556 customers
- No Action: 1291 customers
- Cross-Selling: 474 customers
- No Recommendations: 131 customers

Customers by segment:

- Champions: 1703 customers
- Lost Customers: 1291 customers
- Potential Loyalists: 969 customers
- Loyal Customers: 931 customers
- At Risk: 925 customers

Recommendation coverage:

- Customers with recommendations: 4397 (75.6%)
- Customers without recommendations: 1422
- Average recommendations per customer: 2.9

4. DATA INTEGRITY CHECK:

- Duplicate customers: 0 (should be 0)
- Unique customers: 5819
- Total rows: 5819

5. SAMPLE OF FINAL CUSTOMER 360 RECOMMENDATION TABLE:
Sample customers with recommendations:

```
    customer_id customer_segment  total_spent  total_orders  primary_scenario  \
1       12347.0        Champions      4921.53             8    Revenue Growth
2       12348.0  Loyal Customers      1658.40             5    Revenue Growth
3       12349.0        Champions      3405.99             3    Revenue Growth
6       12352.0        Champions      1459.18             7    Revenue Growth
7       12353.0          At Risk       406.76             2  Churn Prevention
8       12354.0          At Risk      1079.40             1  Churn Prevention
9       12355.0          At Risk       947.61             2  Churn Prevention
10      12356.0        Champions      5611.73             6    Revenue Growth
11      12357.0  Loyal Customers     17437.66             2    Revenue Growth
12      12358.0        Champions      3447.07             5    Revenue Growth

   recommendation_1_product               recommendation_1_name  \
1                    85123A  WHITE HANGING HEART T-LIGHT HOLDER
2                     22423            REGENCY CAKESTAND 3 TIER
3                    85123A  WHITE HANGING HEART T-LIGHT HOLDER
6                    85123A  WHITE HANGING HEART T-LIGHT HOLDER
7                     21754            HOME BUILDING BLOCK WORD
8                    85099B         JUMBO BAG RED WHITE SPOTTY
```

```
9                  85099B        JUMBO BAG RED WHITE SPOTTY 
10                  47566                     PARTY BUNTING
11                 85123A  WHITE HANGING HEART T-LIGHT HOLDER
12                  21755            LOVE BUILDING BLOCK WORD

   recommendation_1_category recommendation_2_product  \
1                 HOME_DECOR                    84879
2        KITCHEN_FOOD_UTENSIL                    47566
3                 HOME_DECOR                    22469
6                 HOME_DECOR                    84879
7          FURNITURE_STORAGE                    23298
8          TEXTILES_CLOTHING                    21754
9          TEXTILES_CLOTHING                    21754
10       KITCHEN_FOOD_UTENSIL                    22139
11                HOME_DECOR                    84879
12            BEAUTY_PERSONAL                    21790

             recommendation_2_name recommendation_2_category  \
1      ASSORTED COLOUR BIRD ORNAMENT                HOME_DECOR
2                    PARTY BUNTING      KITCHEN_FOOD_UTENSIL
3            HEART OF WICKER SMALL                HOME_DECOR
6      ASSORTED COLOUR BIRD ORNAMENT                HOME_DECOR
7                   SPOTTY BUNTING            GARDEN_OUTDOOR
8          HOME BUILDING BLOCK WORD         FURNITURE_STORAGE
9          HOME BUILDING BLOCK WORD         FURNITURE_STORAGE
10  RETRO SPOT TEA SET CERAMIC 11 PC        KITCHEN_FOOD_UTENSIL
11     ASSORTED COLOUR BIRD ORNAMENT                HOME_DECOR
12                VINTAGE SNAP CARDS           BEAUTY_PERSONAL

   recommendation_3_product              recommendation_3_name  \
1                     22469               HEART OF WICKER SMALL
2                     21212     PACK OF 72 RETRO SPOT CAKE CASES
3                     22138        BAKING SET 9 PIECE RETROSPOT 
6                     22469               HEART OF WICKER SMALL
7                     21791  VINTAGE HEADS AND TAILS CARD GAME 
8                     21791  VINTAGE HEADS AND TAILS CARD GAME 
9                     21791  VINTAGE HEADS AND TAILS CARD GAME 
10                    22111         SCOTTIE DOG HOT WATER BOTTLE
11                    22469               HEART OF WICKER SMALL
12                    21915              RED  HARMONICA IN BOX 

   recommendation_3_category
1                 HOME_DECOR
2       KITCHEN_FOOD_UTENSIL
3                 HOME_DECOR
6                 HOME_DECOR
7                 TOYS_GAMES
8                 TOYS_GAMES
```

```
9                TOYS_GAMES
10      KITCHEN_FOOD_UTENSIL
11                HOME_DECOR
12            BEAUTY_PERSONAL


Final Customer 360 Recommendation Table completed!

6. EXPORTING RECOMMENDATION TABLE:
   • Exported to: datasets/customer_360_recommendations.csv
```

7 Summary and Recommendation

7.1 Key Insights from Analysis

- **Revenue Concentration**
  - The business is **heavily reliant on a small group of high-value customers**. Champions represent **29% of the base** but deliver **79% of revenue**.
  - This imbalance means retention of Champions is mission-critical. Even a small drop in their activity would have a disproportionate financial impact.
  - It also highlights the challenge: the broader base contributes far less but they are key for business growth.
- **Retention Gaps**
  - The **Potential Loyalists (17%)** and **At Risk (16%)** segments together account for a third of the customer base. They are the **"pivot group"**: with proper engagement, they can graduate into Loyal or even Champions, but without intervention, many will slide into Lost.
  - **Lost Customers (22%)** are numerous but low-value, contributing only **2% of revenue**. This confirms that high-cost reactivation is not efficient here. Instead, automated or seasonal touchpoints are sufficient.
- **Churn Prediction (90-day window)**
  - Our churn model reveals a **25.3% churn rate among active customers**, with **1,556 predicted as high risk (>80% probability)**. This represents **$1.5M in revenue at risk** if left unaddressed.
  - Drivers of churn are clear: **extended purchase gaps and limited category diversity**. Customers who fail to buy regularly or who remain concentrated in just one or two categories are much more likely to leave.
  - Conversely, customers with **consistent frequency and multi-category engagement** are significantly "stickier" and more resilient to churn.
- **Category Stickiness**
  - A clear split emerges between **revenue categories** and **retention categories**.
    * **Home Decor & Kitchen** drive the highest sales but are weak in retention, many customers churn if they stay confined here.

    * **Christmas, Furniture, and Toys** show much stronger retention (>60%) and are disproportionately favored by Champions.

  – Even more critical is **category diversification**. CLV grows almost **10x when customers purchase from 9 categories vs 1**. This proves that true loyalty is built not just on repeat purchases, but on broadening the relationship across categories.

7.2 Pilot Plan – Retention Campaign

From the insights that have been gathered, the pilot plan strategy for retention campaign is as below

1. **Protect & Reward Champions**
   - Champions are the backbone of the business and must be shielded.
   - Offer **VIP perks, early access, exclusive bundles, and personalized product recommendations** in their favorite categories.
   - The goal is not to change behavior but to reinforce loyalty and reduce churn risk.
2. **Upgrade Loyal Customers → Champions**
   - Loyal customers already show consistency and are primed to move up.
   - **Targeted upselling in Textiles and Furniture** plus cross-sell into sticky categories can accelerate their growth.
   - Use **personalized offers to shorten purchase cycles** and push them closer to Champion-level frequency.
3. **Nurture Potential Loyalists**
   - This group is the largest opportunity. They are active but not yet committed.
   - Launch **cross-selling campaigns** that introduce them to new categories.
   - Use **loyalty points, curated bundles, or "complete the set" offers** to encourage repeat activity and diversify baskets.
4. **Recover At Risk**
   - These customers are disengaging, but not lost yet.
   - Use **win-back campaigns tied to sticky categories** (Furniture, Toys, Christmas) which have historically reactivated customers.
   - Leverage churn scores to **prioritize high-value At Risk customers first** to maximize ROI.
5. **Lost Customers**
   - Their revenue contribution is minimal.
   - Keep **re-engagement low-cost and automated** (e.g., seasonal promotions, newsletters). Heavy investment here would dilute focus from higher-value opportunities.

7.2.1 Personalized Recommendation Scenarios

As a tactical initiatives for the retention campaign, here is the proposed approach

We will use 3 Scenario Recommendation

1. **Drive Revenue (Champions & Loyal)**

- Recommend the **top 3 products in a customer's favorite category** that they have not yet purchased.
- Focus on **Home Decor & Kitchen**, as they dominate spending and generate quick wins for revenue lift.

2. **Cross-Selling for Stickiness (Potential Loyalists & At Risk)**
   - Recommend **1 product from 3 related categories** (based on correlation analysis).
   - This diversifies the basket, pushing customers into the **5–9 category range** where CLV is significantly higher.

3. **Churn Prevention (High Risk)**
   - Target the **1,556 high-risk customers** flagged by the churn model.
   - Recommend **products from sticky categories (Furniture, Toys, Textiles)**, combined with promotional offers (e.g., **25% discount + free shipping**).
   - The goal is to reactivate them with categories that historically drive retention.