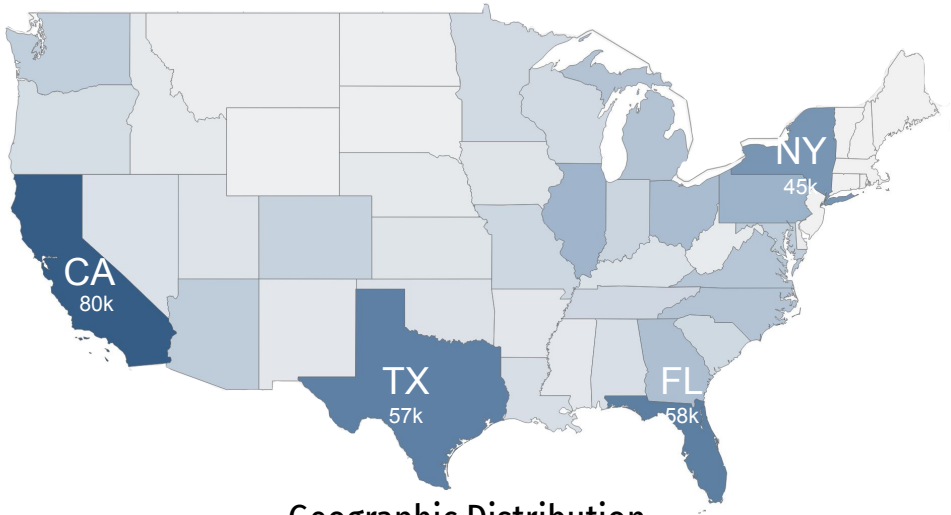
Customer Segmentation Analysis

Contents

Exploratory Data Analysis

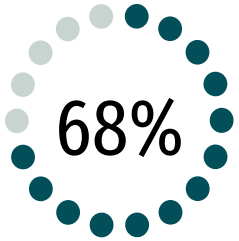
Diversified customer composition



Geographic Distribution

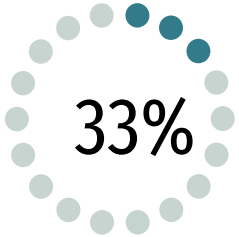
(According to counts of zip codes)

Major markets spread across different regions of the country



Devices for Home Usage

Max=Home-Personal Min=Office



Users with Household of 2

Min=9 Max=2

Feature Engineering

01

Outlier and Null Value Treatment

- Dropped outliers of each column
- Replaced nulls with column means, randomly assigned values based on original ratios

02

Calculation and Variable Pre-Selection

- Calculated '# of printers' and RFM (Recency, Frequency, Monetary) value for each customer
- Dropped variables unrelated to the current problem, Feature selection

03

Categorical Variable Encoding

- Transformed some of the categorical variables to dummy values

04

Scaling

- As we are using the k-means clustering model, which groups the customers according to the distance between them, we scaled the data to avoid the influence of units

“RFM” Analysis

1

Recency

Time since the most recent purchase

2

Frequency

Average time between purchases (Intensity)

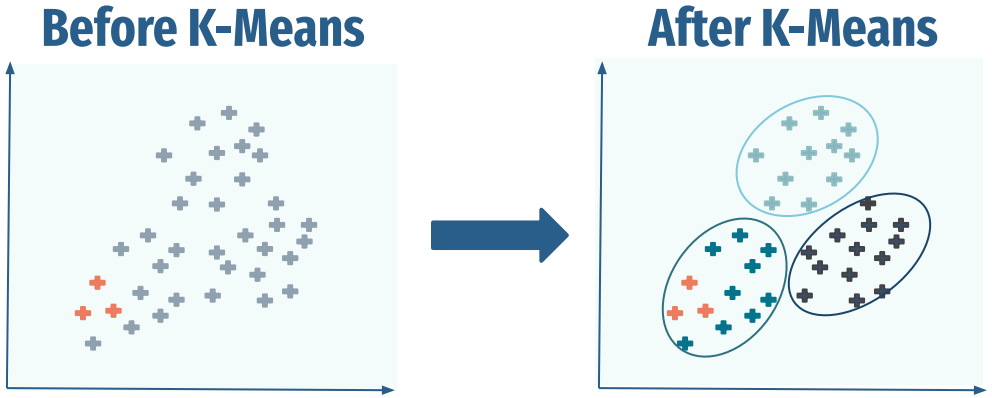
3

Monetary

Total and average purchase values

Model Introduction K-Means Clustering

- Customer RFM
- Customer Geographics
- Customer Interests
- Customer Usage



K-Means clustering will group together customers with similar characteristics

Has the clustering worked ?

To validate the clustering algorithm, we checked for significant differences in **proportions of repurchase customers** among the groups.

Model Evaluation

Metrics

We used the elbow method to determine the optimal number of clusters (k) for our clustering algorithm.

Result

The “elbow curve” describes the “inertia”, which is the sum of squared distances between each data point and its nearest centroid, against the number of clusters (k).

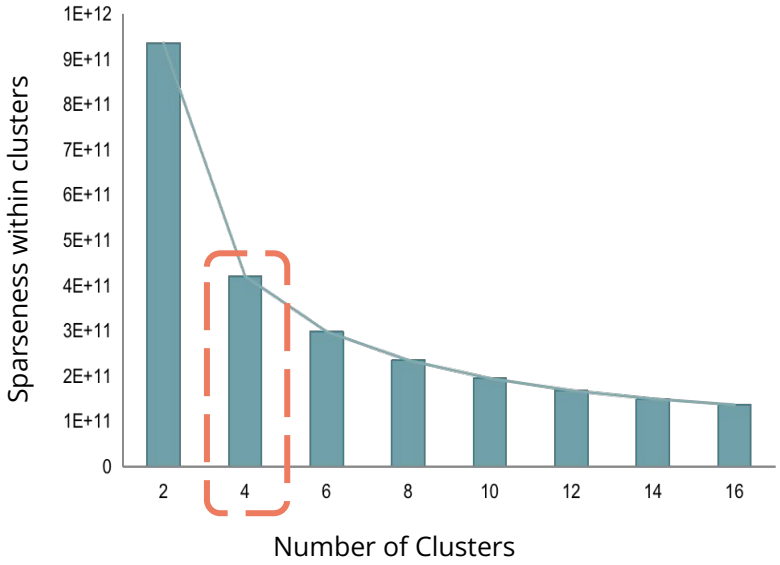
Decision

Our evaluation result indicates the optimal k = 4 after which the WCSS (Within Cluster Sum of Square) decreases slowly.

Contents

Model Result & Evaluation

Choosing the number of clusters



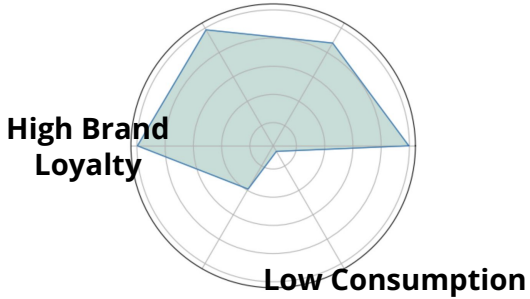
Optimal Number of Clusters = 4

Data of each cluster

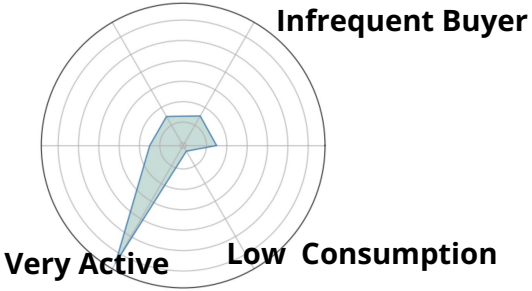
Cluster	Repurchased	# of Printers	Ink/Mo	Active Days
1	No	1	x cc	xxx
2	No	1	x cc	xxx
3	No	1	x cc	xxx
4	Yes	2	xx cc	xxx

Cluster 4 contains the most of the repurchasing customers

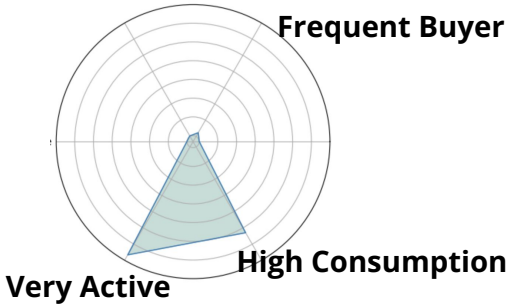
Cluster 1
Casual Users



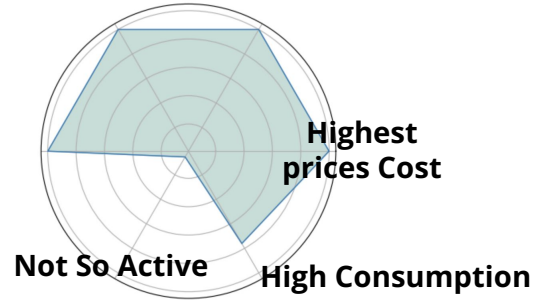
Cluster 2
Exploratory Users



Cluster 3
Heavy Users



Cluster 4
Loyal Users



Note: Graphs created with Google Cloud products

Recommendations

Different Clusters, Different Strategies

01 Purchase

- Frequent buyers: promote either cheaper or more pricey products

- Infrequent buyers: promote medium-priced products

02 Printing

- Diversified habits require more flexible usage plans

- Leverage Smart data that captures usage patterns to design more ink subscription options

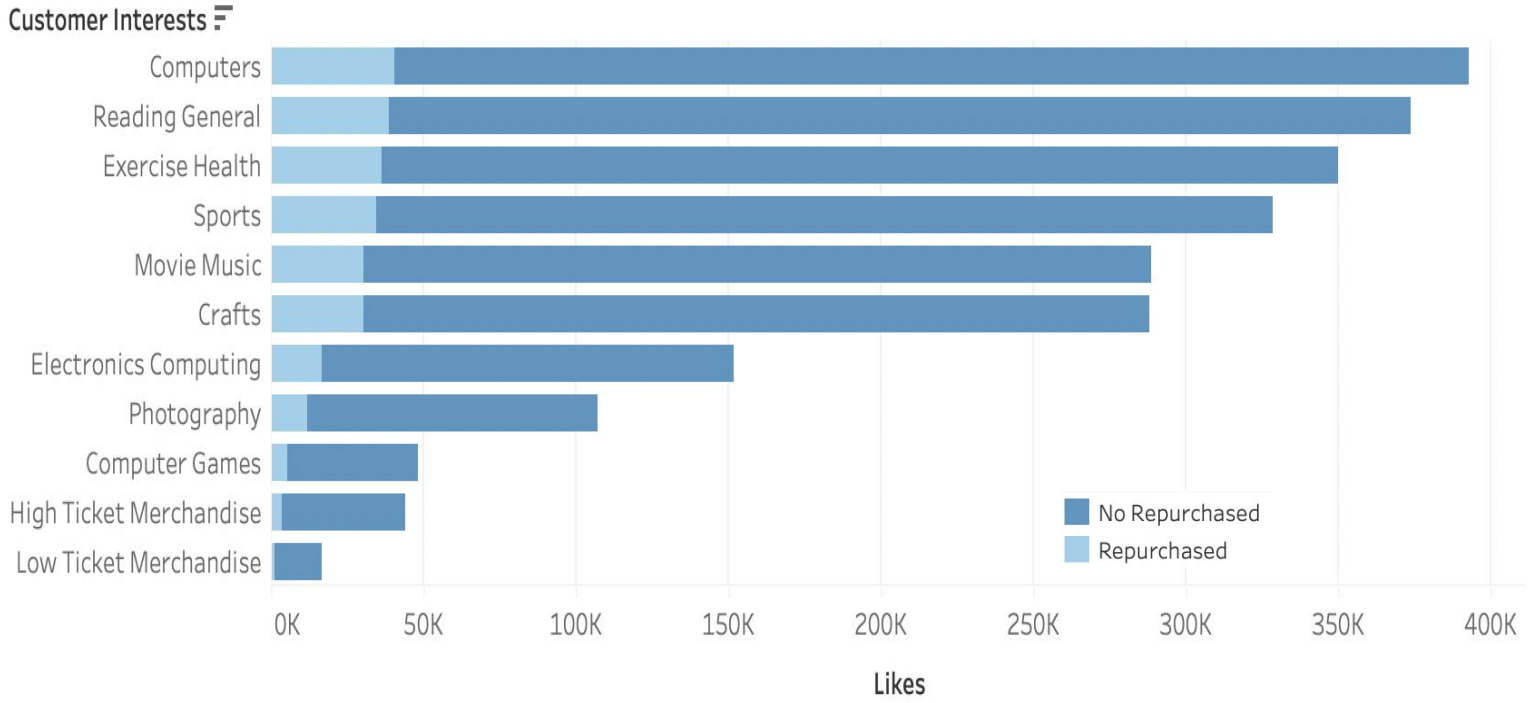
03 Retention

- Improve activeness might not improve loyalty

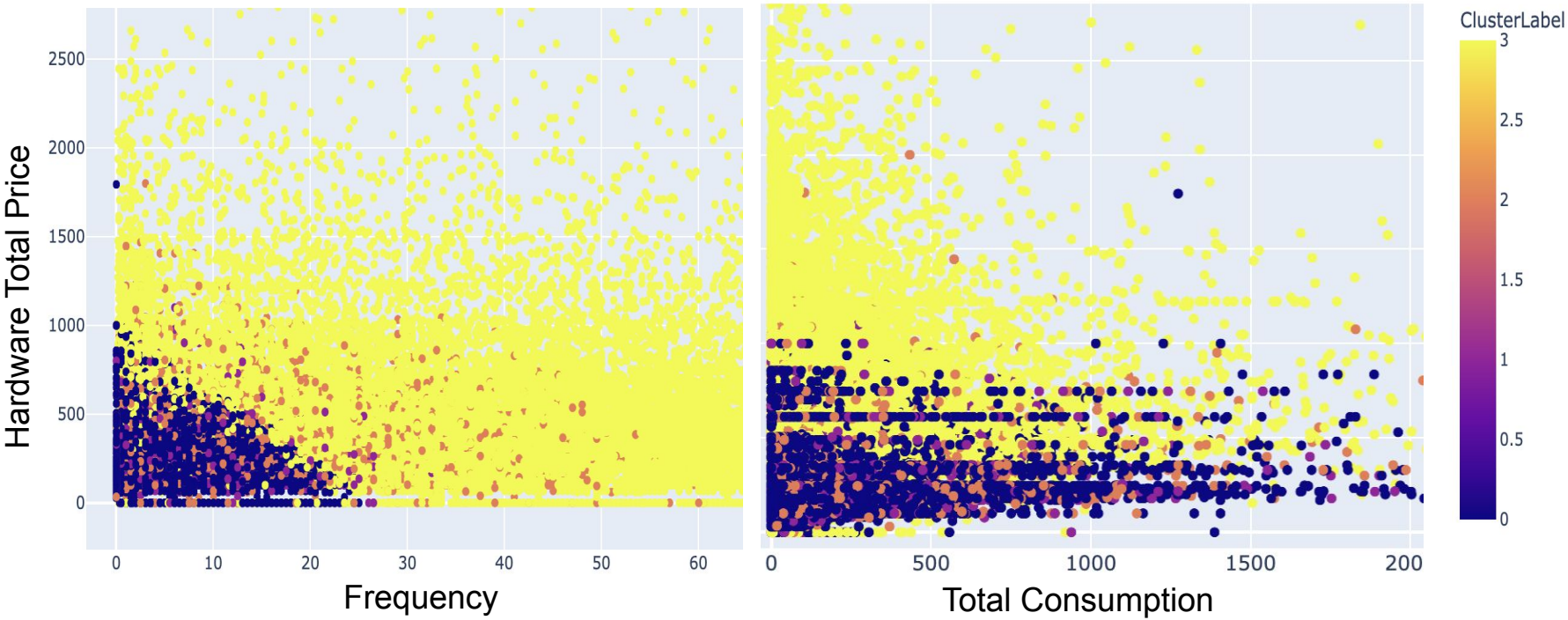
- Usage plan subscription increases stickiness while bringing stable revenue

Appendix

Customer Interests — Repurchased VS. No Repurchase



Frequency / Consumption vs Hardware Total Price



Model Evaluation - ARI Metric

Metrics

We used the Adjusted Rand Index is clustering evaluation metric that measures the similarity between the true labels and the predicted cluster labels.

Result

We get Adjusted Rand Index = 0.05900098600331566

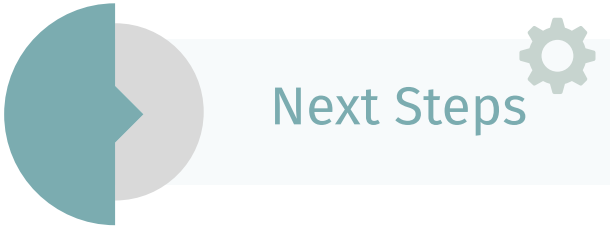
Decision

The value of ARI is close to zero, which indicates that the true labels and the predicted cluster labels is no better than what would be expected by chance

Discussion

Limitation and next steps

1. Detailed **product data** and **customer preference data** not available
2. No **synergy between products** is considered
3. K-Means clustering may **has less prediction power for future new data**



Predict the repurchase **propensity score** for each customer to better quantify the clustering results