greatlearning

Project Summary

Batch details	PGP-DSE-FT-BLR-AUG-2023
Team members	1. Abhishek Khanra 2. Akshen Dhami 3. Harshal Hinge 4. Ritesh Bhandari 5. Yash Rajendra Jadhav
Domain of Project	Retail
Proposed project title	Predicting Customer Acquisition Cost for a store
Group Number	1
Team Leader	Abhishek Khanra
Mentor Name	Ms. Prachi Tare

Date: 27.02.2024

ABHISHEK KHANRA

Signature of the Mentor   Signature of the Team Leader

Table of Contents

Summary of Problem statement, data and findings

The capstone project aims to predict cost of acquisition of customers for CFM(Convenient Food Mart), a chain of convenient stores in the United States of America, based on the demographic factors of customers, types of stores and products among others using different media channels(newspaper, radio, TV, etc.)

Proposed solution to the problem vis-à-vis findings of the project

a. Predicting customer acquisition cost using the features (customer related, product/service related, store related) in the dataset using Decision Tree Regressor

b. Finding the most important factors impacting customer acquisition cost, which could be recommended to the company for prioritizing for reducing customer acquisition cost.

c. Clustering via STP Principle (segmentation, targeting and positioning)-segment customers and then accordingly target and position the company towards customers with lowest customer acquisition cost or highest avg. yearly income or a combination of both etc.)

Findings:

a. After using various models (Linear Regression, Decision Tree Regressor, Random Forest Regressor, Ada Boost Regressor, Gradient Boost Regressor, XG Boost Regressor, Stacking Regressor (1. base models-[Random Forest Regressor, XG Boost Regressor],final estimator-Decision Tree Regressor. 2. Base models-[Decision Tree Regressor, Random Forest Regressor],final estimator-XG Boost Regressor) for different selection of features, final model selected having best mixture of low variance and low bias error was XG Boost Regressor model with 16 features(15 independent features, 1 dependent feature). Train MSE-2661(RMSE-51.58), Test MSE-3154(RMSE-56.16), Cross Validation Score MSE-16761(RMSE-129.46).
b. Most important factors impacting customer acquisition cost for our best model XG Boost Regressor for 16 features:
   a. Video Store (presence or absence)
   b. Media Type (media channels)
   c. Sales District (store district)
   d. Promotion Period
   e. Store sqft (area of store)
   f. Promotion Name (Types of promotion)
   g. Store Type (Types of store)

c. K Prototype Clustering (since our final model contained a mixture of numerical and categorical features). 2 clusters:

1. Cluster 0 - price discount searchers buying costlier products and more number of products.

2. Cluster 1 - sale day searchers buying cheaper products and less number of products.

Implications:

a. To reduce customer acquisition cost, we should focus on reducing promotion period to get ideal promotion period, while deciding varying customer acquisition cost based on different type of store, promotion type, media type, location of store and store area.
b. Attempt 2 different strategies for acquiring customer [1. Price discount buying more than 3 units of product, average SRP of USD 3 and unit cost of USD 4 (discount-1.05 USD) and net weight of product >12 ounces. 2. Sale day buying less than 3 units of product, average SRP < USD 3 and unit cost <USD 3 (discount -0.09 USD) and net weight of product < 12 ounces.

Overview of the final process

Dataset and Domain

- Data Dictionary

1. Product

S No.	Feature	Description
1.	product_class_id	ID of product_class
2.	product_id	ID of products
3.	brand_name	Brand names of products
4.	product_name	Name of products
5.	SKU	Stock Keeping Unit(SKU) – scannable bar code (alphanumeric combination that tracks price, product details, manufacturer, and point-of-sale).
6.	SRP	Suggested Retail Price/Manufacturer's suggested Retail Price/Recommended Retail Price/List Price- price at which its manufacturer notionally recommends that a retailer sell the product.
7.	gross_weight	Weight in ounces (1 fluid ounce = 29.57 ml, 1 ounce =28.34 g)
8.	net_weight	Weight in ounces (1 fluid ounce = 29.57 ml, 1 ounce =28.34 g)
9.	recyclable_package	No-0, Yes-1
10.	low_fat	No -0, Yes-1
11.	units_per_case	Units of products in 1 case
12.	cases_per_pallet	Number of cases per pallet (pallet is flat transport structure, which supports goods in a stable fashion while being lifted by a forklift, a

		pallet jack, a front loader, a jacking device, or an erect crane).
13.	shelf _width	Width of product shelf(in inches)
14.	shelf_height	Height of product shelf(in inches)
15.	shelf_depth	Depth of product shelf([distance between 2 shelf] in inches)

2. Customer

S No.	Feature	Description
1.	customer_id	ID of customer
2.	account_num	Account_number of customer in transaction
3.	l_name	Last name of customer
4.	f_name	First name of customer
5.	Mi	Middle initial of customer
6.	Address1	First part of address of customer
7.	Address2	Second part of address of customer
8.	Address3	Third part of address of customer
9.	Address4	Fourth part of address of customer
10.	City	City of customer
11.	State_province	State of customer
12.	Postal_code	Postal_code of customer
13.	Country	Country of customer
14.	Customer_region_id	Region id of customer
15.	phone_1	First number of customer
16.	Phone_2	Second number of customer
17.	Birthdate	Date of birth of customer
18.	Marital_status	Marital status of customer (‘M’-Married, ”S”-Single)
19.	Yearly_income	Yearly income of customer ('$30K - $50K', '$70K - $90K', '$50K - $70K', '$10K - $30K', '$90K - $110K', '$110K - $130K', '$130K - $150K', '$150K +')
20.	Gender	Gender of customer (“F”- Female, “M”- Male)
21.	Total_children	Total number of children of customer[0,1,2,3,4,5]
22.	num_children_at_home	Number of children at home(who haven’t accompanied the customer) [0,1,2,3,4,5]
23.	education	Highest education completed by customer ['Partial High School', 'Bachelors Degree', 'Partial College', 'High School Degree', 'Graduate Degree']
24.	date_accnt_opened	Date on which customer opened an account with the store
25.	Member_card	Type of store membership card owned by customer (“Normal”,” Bronze”, “Silver”, “Gold”]

26.	occupation	Occupation of customer ['Skilled Manual', 'Professional', 'Manual', 'Management','Clerical']
27.	Houseowner	Whether a customer owns a house or not [“Y”,”N”]
28.	Num_cars_owned	Number of cars owned by a customer [0,1,2,3,4,]
29.	Fullname	Full name of customer

3. Product_class

S No.	Feature	Description
1.	product_class_id	ID of product_class
2.	subcategory	Subcategory of product_class
3.	category	category of product_class
4.	department	Department comprising of similar category of product_class (group of categories that meet related needs)
5.	family	Product family ['Food', 'Drink', 'Non-Consumable']

4. Promotion

S No.	Feature	Description
1.	promotion_id	ID of promotion
2.	Promotion_district_id	District id of promotion
3.	promotion_name	Name of promotion
4.	Media_type	Channels of promotion ['Product Attachment', 'Radio', 'In-Store Coupon', 'Sunday Paper, Radio', 'Daily Paper', 'Sunday Paper, Radio, TV', 'Daily Paper, Radio, TV', 'TV', 'Cash Register Handout','Street Handout', 'Daily Paper, Radio', 'Sunday Paper','Bulk Mail']
5.	cost	Customer acquisition cost spent on promotion (in USD)
6.	Start_date	Start_date of promotion
7.	End_date	End_date of promotion

5. Region

S No.	Feature	Description
1.	region_id	ID of region
2.	Sales_city	City of sales
3.	Sales_state_province	State of sales
4.	Sales_district	District of sales

5.	Sales_region	Region of sales ['No Region', 'Central West', 'Mexico Central', 'South West', 'Mexico West', 'Canada West', 'North West', 'Mexico South']
6.	Sales_country	Country of Sales ['No Country', 'USA', 'Mexico', 'Canada']
7.	Sales_district_id	District_id of sales

6. Sales

S No.	Feature	Description
1.	region_id	ID of region
2.	Sales_city	City of sales
3.	Sales_state_province	State of sales
4.	Sales_district	District of sales
5.	Sales_region	Region of sales ['No Region', 'Central West', 'Mexico Central', 'South West', 'Mexico West', 'Canada West', 'North West', 'Mexico South']
6.	Sales_country	Country of Sales ['No Country', 'USA', 'Mexico', 'Canada']
7.	Sales_district_id	District_id of sales

7. Time_by_day

S No.	Feature	Description
1.	time_id	ID of transaction time
2.	The_date	Date of transaction (YYYY-MM-DD)
3.	The_day	Day of transaction [‘Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday','Sunday']
4.	The_year	year of sales [1996,1997,1998,1999]
5.	The_month	month of sales [1,2,3,4,5,6,7,8,9,10,11,12]
6.	Day_of_month	Day of month of Sales [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31]
7.	Day_since_epoch	No of days since epoch date(1900-01-01). In a computing context, an epoch is the date and time relative to which a computer's clock and timestamp values are determined. The epoch traditionally corresponds to 0 hours, 0 minutes, and 0 seconds (00:00:00) Coordinated Universal Time (UTC) on a specific date, which varies from system to system. Apple macOS considers its Epoch Time as starting from January 1, 1904

8. Store

S No.	Feature	Description

1.	store_id	ID of store
2.	Store_type	Type of store ['HeadQuarters', 'Supermarket', 'Small Grocery','Gourmet Supermarket', 'Deluxe Supermarket', 'Mid-Size Grocery']
3.	Region_id	Region_id of store
4.	Store_name	Name of store
5.	Store_number	Number of store
6.	Store_street_address	Street address of store
7.	Store_city	city of store
8.	Store_state	State of store
9.	Store_postal_code	Postal code of store
10.	Store_country	Country of store [USA, Mexico, Canada]
11.	Store_manager	Name of store manager
12.	Store_phone	Phone number of store
13.	Store_fax	Fax of store
14.	First_opened_date	Datetime of opening of store
15.	Last_remodel_date	Datetime of last remodeling of store
16.	Store_sqft	Total store area (in square feet)
17.	Grocery_sqft	Grocery product portion of store area (in square feet)
18.	Frozen_sqft	Frozen product portion of store area (in square feet)
19.	Meat_sqft	Meat product portion of store area (in square feet)
20.	Coffee_bar	Whether store has coffee bar or not (0- No, 1- Yes)
21.	Video_store	Whether store has video_store or not (0- No, 1- Yes)
22.	Salad_bar	Whether store has salad bar or not (0- No, 1- Yes)
23.	Prepared_food	Whether store has cooked/prepared food or not (0- No, 1- Yes)
24.	florist	Whether store has florist or not (0- No, 1- Yes)

- Variable categorization (count of numeric and categorical)

  Numerical – 21
  ['store_sales', 'store_cost', 'unit_sales', 'SRP', 'gross_weight', 'net_weight', 'units_per_case', 'cases_per_pallet', 'shelf_width','shelf_height', 'shelf_depth', 'cost', 'store_sqft', 'grocery_sqft','frozen_sqft', 'meat_sqft', 'total_children', 'num_children_at_home','num_cars_owned', 'day_since_epoch', 'promotion_period']

  Categorical – 73
  ['product_id', 'time_id', 'customer_id', 'promotion_id', 'store_id','product_class_id', 'brand_name', 'product_name', 'SKU','recyclable_package', 'low_fat',

'subcategory', 'category','department', 'family', 'promotion_district_id', 'promotion_name',
'media_type', 'start_date', 'end_date', 'region_id', 'sales_city','sales_state_province', 'sales_district', 'sales_region','sales_country', 'sales_district_id', 'store_type', 'store_name','store_number', 'store_street_address', 'store_city', 'store_state',
'store_postal_code', 'store_country', 'store_manager', 'store_phone',
'store_fax', 'first_opened_date', 'last_remodel_date', 'coffee_bar',
'video_store', 'salad_bar', 'prepared_food', 'florist', 'account_num','lname', 'fname', 'mi', 'address1', 'address2', 'city','state_province', 'postal_code', 'country', 'customer_region_id','phone1', 'phone2', 'birthdate', 'marital_status', 'yearly_income',
'gender', 'education', 'date_accnt_opened', 'member_card', 'occupation',
'houseowner', 'fullname', 'the_date', 'the_day','the_year','the_month','day_of_month']

- Pre Processing Data Analysis (count of missing/ null values, redundant columns, etc.)

  Null values:

	columns with null values	number of null values	% of null values
12	address2	258116	95.697761
0	promotion_id	198969	73.768723
1	promotion_district_id	198969	73.768723
2	promotion_name	198969	73.768723
3	media_type	198969	73.768723
4	cost	198969	73.768723
5	start_date	198969	73.768723
6	end_date	198969	73.768723
13	promotion_period	198969	73.768723
11	mi	113978	42.257897
7	store_sqft	38811	14.389367
8	grocery_sqft	38811	14.389367
9	frozen_sqft	38811	14.389367
10	meat_sqft	38811	14.389367

- Null value treatment – drop rows for subset="cost" [target variable] and thereafter median imputation (for skewed columns).

- Redundant columns
- (columns [store_city,sales_city],[sales_state_province,store_state],[sales_country,store_country,country],[store_postal_code (only 1 unique value), add3 and add4 (100% null values)],[also there are other columns which represent the same thing, for eg. [Promotion_name, promotion_id],[store_name, store_id],[product_name, product_id], etc.)

Data Exploration (EDA)

- Relationship between variables
  - multi-collinearity – not observed as per heatmap but after doing Linear Regression (condition number was very high (e^{21})). So high multicollinearity.

  - distribution of variables – skewness, kurtosis
  - Skewness
    units_per_case,shelf_height,total_children,num_cars_owned, days_since_epoch are negatively skewed(left skewed). rest all features are positively skewed.
  - Kurtosis
    store_sales,store_cost,frozen_sqft,meat_sqft,num_children_at_home and promotion_period are leptokurtic. Rest all columns are platykurtic.

  - presence of outliers and its treatment

  - outliers via IQR method is observed in store_sales,store_cost,unit_sales, frozen_sqft,meat_sqft,num_children_at_home and promotion_period.

    Outlier treatment was not done.

  - statistical significance of variables

    Numerical columns vs target column ("cost") – mannwhitneyu test (skewed columns)

    All columns were significant.

    Categorical columns vs target column ("cost") – mannwhitneyu/kruscal-wallis test (skewed column)

    9 columns out of 73 columns were insignificant ["product_id","product_class_id","brand_name","product_name", "SKU", "subcategory","category","department","family"]

Feature Engineering

- Whether any transformations required – Decision Tree Regressor and other ensemble learning regressor methods which are not distance based algorithms. So scaling is not required.
- 1 feature was created [Promotion Period= last_date - first_date]
- Feature selection
- 1. 56 columns – dropping id columns, address, date columns
- 2. 43 columns – dropping a lot of name columns (customer_name, product_name, etc.)
- 3. 16 columns

greatlearning

- 4. 21 columns

Algorithms:

1. Linear Regression
2. Decision Tree Regressor
3. Random Forest Regressor (Bagging Technique)
4. Ada Boost Regressor (Boosting Technique)
5. Gradient Boost Regressor
6. XG Boost Regressor,
7. Stacking Regressor (1. base models-[Random Forest Regressor, XG Boost Regressor],final estimator-Decision Tree Regressor. 2. Base models-[Decision Tree Regressor, Random Forest Regressor],final estimator-XG Boost Regressor)
8. PCA
9. K-Prototype Clustering

Combination Techniques:

1. Reducing subcategories of categorical columns (high cardinality) with label encoding.
2. PCA with only numerical columns and PCA + 9 categorical columns (from our best model – XGB-16 columns)
3. Building models with best 31 columns (from 43) and 16 columns (from 31) using Sequential Feature Selector (SFS(linear regression)-mlxtend library), 21 columns (from 43) using SFS (XGBoostRegressor)-mlxtend library).
4. Building models with best 31 columns (from 43) and 16 columns (from 31) using best features from relative importances of XG Boost Regressor)

Step-by-step walk through of the solution

1. After data preprocessing, 56 columns were selected (based on domain) and all 7 models were run for one-hot encoding, dummy encoding and label encoding. The same was repeated for 43 columns (based on domain).
2. After selecting label encoding, 3 methods of feature selection were used. 31 columns were selected based on domain, SFS (Linear Regression, k_features=”best”) used on a dataset with 56 columns and best columns based on relative importances of best performing model (XGB Regressor) used on a dataset of 43 columns.
3. The same procedure was repeated for selecting 16 columns for domain and for SFS (k_features=”best” for linear regression on 31 columns dataset resulted in 19 independent columns . SFS was used on that dataset for linear regression from which 15 columns were selected (k_features=15)as cost column is target column) and relative importance were calculated on a dataset of 31 columns for XGB Regressor.
4. 21 columns were selected based on domain and SFS (XGB Regressor) used on a dataset with 43 columns.
5. All the models were evaluated using Mean Squared Error (MSE) for train and test data as well as cross-validation score (no of folds=3) for finding the best model with low bias and variance errors.
6. For reducing cardinality of categorical columns, similar subcategories were grouped together to reduce the number of subcategories [department, promotion name] and

sales region was selected instead of sales district. The same procedure was repeated for 43, 31, 16 and 21 columns. Thereafter model evaluation was conducted which gave higher bias error and lower variance error as compared to previous technique.

7. PCA (Principal Component Analysis) was conducted on all 20 independent numerical columns (excluding target column cost) and all 7 models were run and performance was evaluated. PCA components (20) along with 9 categorical columns out of 15 independent columns from XGB model – domain were selected and all 7 models were run and performance was evaluated. Higher bias error and lower variance error was reported as compared to the previous technique.
8. K-prototype clustering for mixed data (9 categorical and 6 numerical features from XGB-16 model- domain) was conducted. Categorical index was given as input for cluster while numerical columns were scaled (standard scaler) and silhouette score was calculated for finding the ideal number of clusters for a range of 2 to 8 clusters (excluding 8). Since highest silhouette score was highest for number of clusters being 2, the same was used to divide customers into 2 clusters.
9. Cluster centroids were calculated and visualized using heatmap to define clusters, However, since the numerical data contained a few discrete variables and encoded categorical columns, insights weren't observed.
10. 5 point summary of both clusters and data visualization of numerical and independent features using clusters as hue (differentiator) provided similarities and differences of customers which provided cluster definition for both clusters.

Model Evaluation

1. Final model was XGB Regressor model based on dataset of 16 features (15 independent, 1 dependent) out of 94 columns based on domain knowledge (selecting important customer, location, product, promotion and store related features).
2. Out of all models built, final model had one of the lowest bias error (train MSE-2661(RMSE -51.58) , test MSE – 3154 (RMSE-56.16)) with the lowest variance error (std (MSE) – 16761). Final model had the best bias-variance tradeoff which results in lowest bias(in this case one of the lowest bias) error and lowest variance (in this case lowest variance) error. Thus, the final model is an effective (low bias error) and generalized (lowest variance error) which works well with new and unseen data.
3. For clustering model, highest silhouette score was 0.209 for no of clusters(k) =2 (for k=2,3,4,5,6,7). The customers were divided into a. price discount searchers buying costlier products and a greater number of products, b. sale day searchers buying cheaper products and a smaller number of products.

Comparison to benchmark

1. The initial benchmark for predicting customer acquisition cost was 10 % difference in train and test RMSE and lowest possible error. The model reported 8.87% difference in test (56.16) and train RMSE (51.58). Although the error itself is far from the lowest possible error (test and train MSE (0)) but this is close enough as the model is a

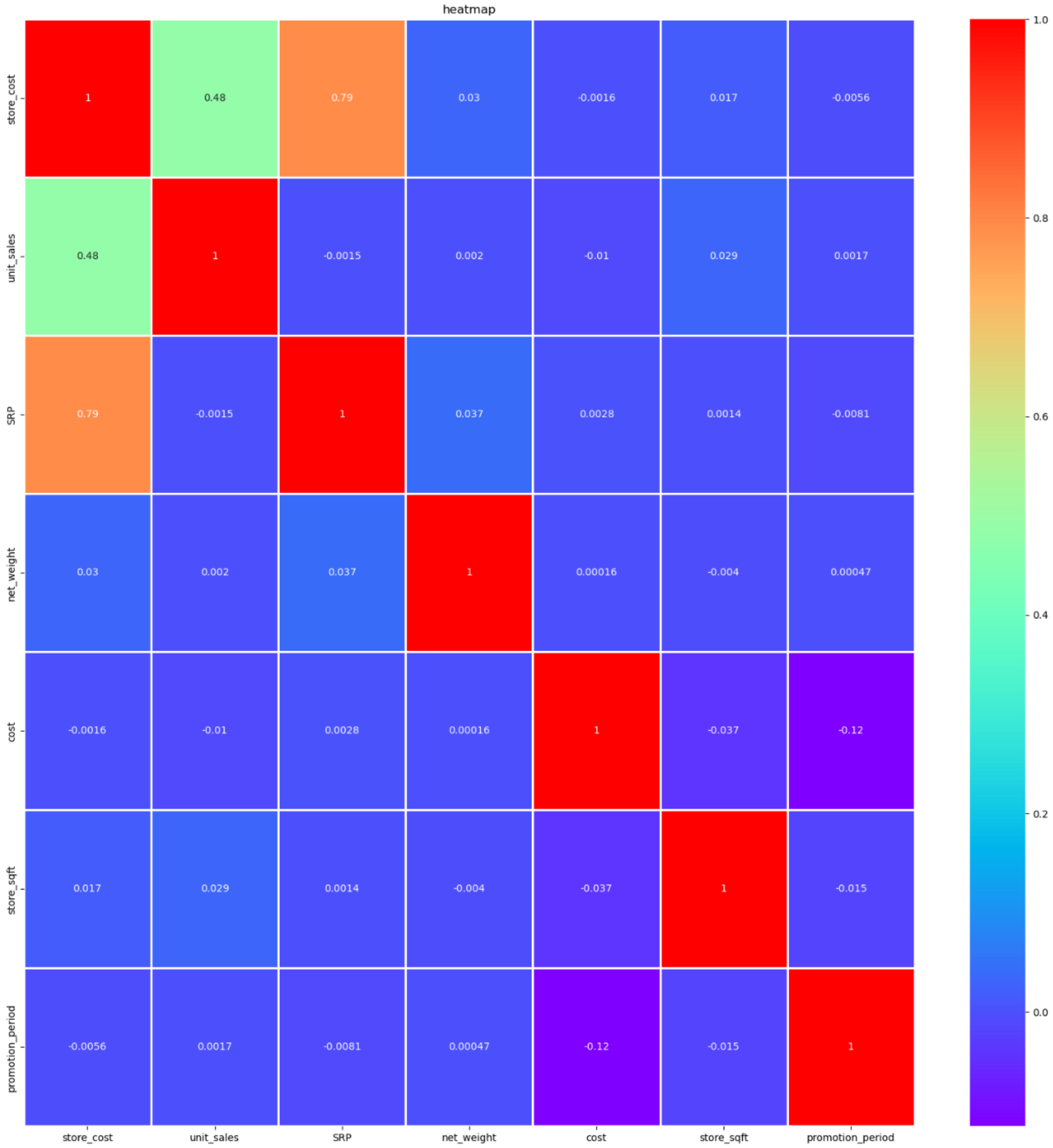
good generalized model (low variance error) and good performing model (low bias error).

2. The initial benchmark for clustering was to make defined clusters. The resultant clusters had a few similar characteristics and a few differentiating characteristics for features which resulted in 2 solidly defined clusters. While not being entirely distinct from each other, the clusters seem to reflect customers who in real life might share a or have a preference of a few characteristics but will also differentiate themselves with others.

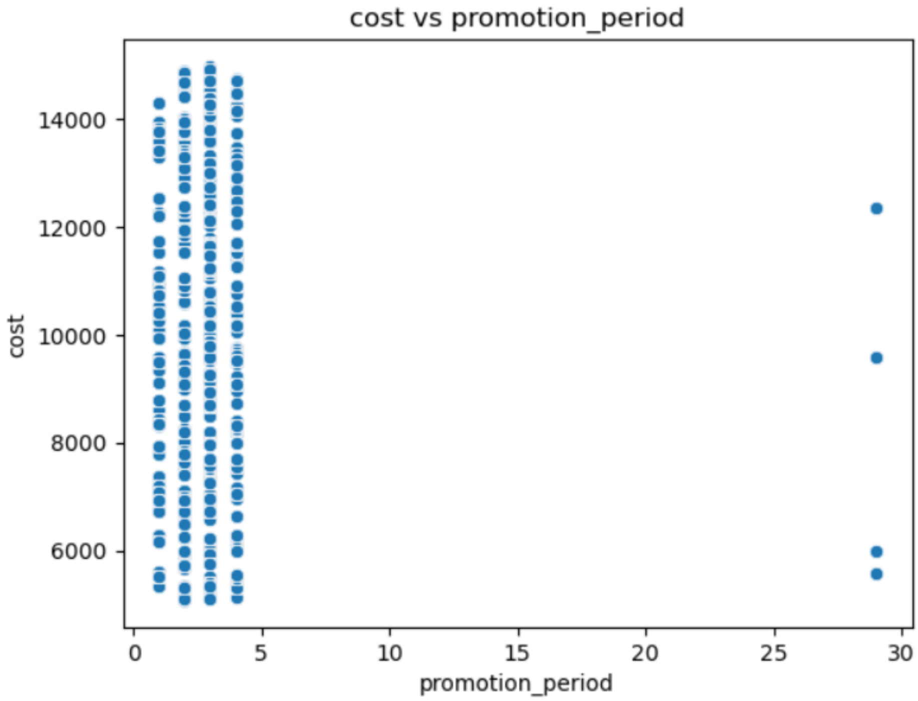
Visualizations

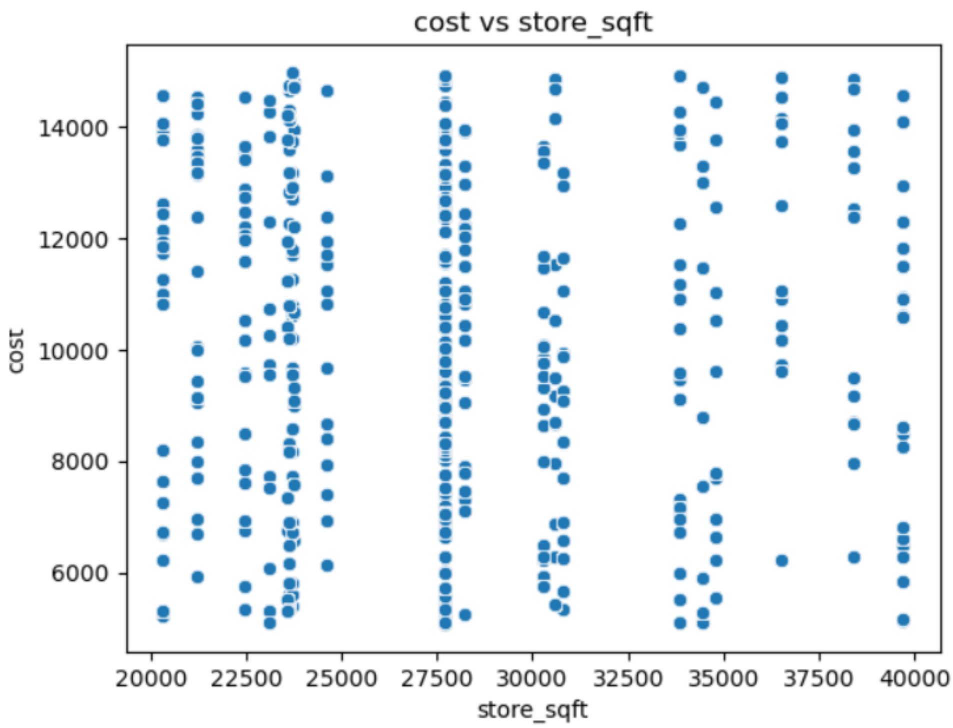
1. Model Estimation
   a. Numerical vs Target (Cost)

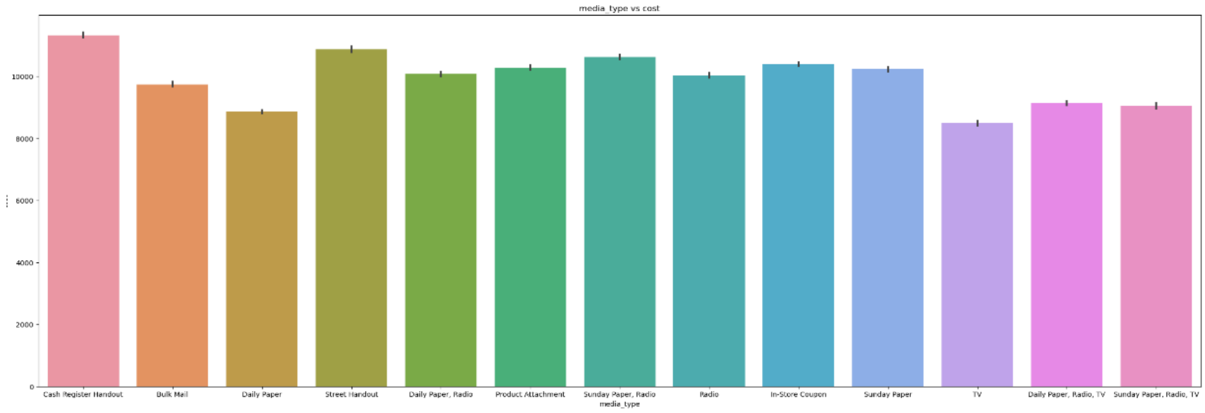
heatmap

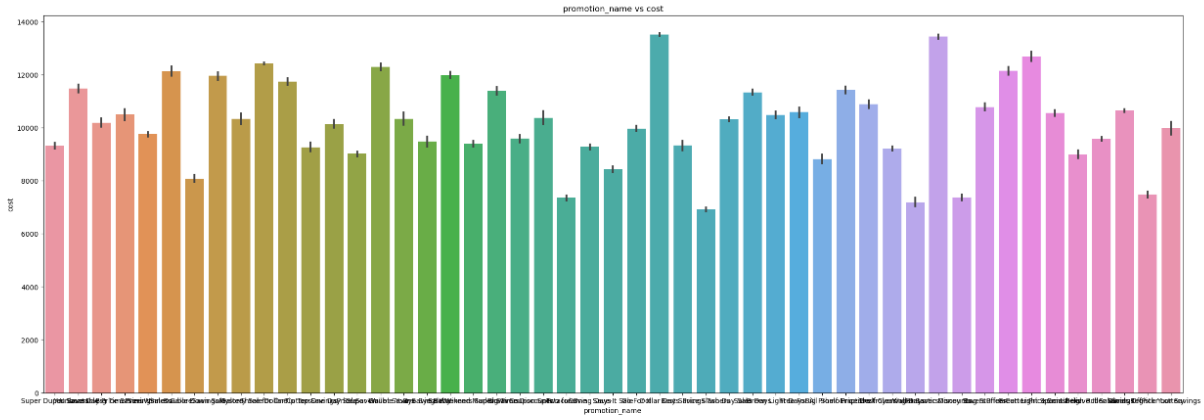


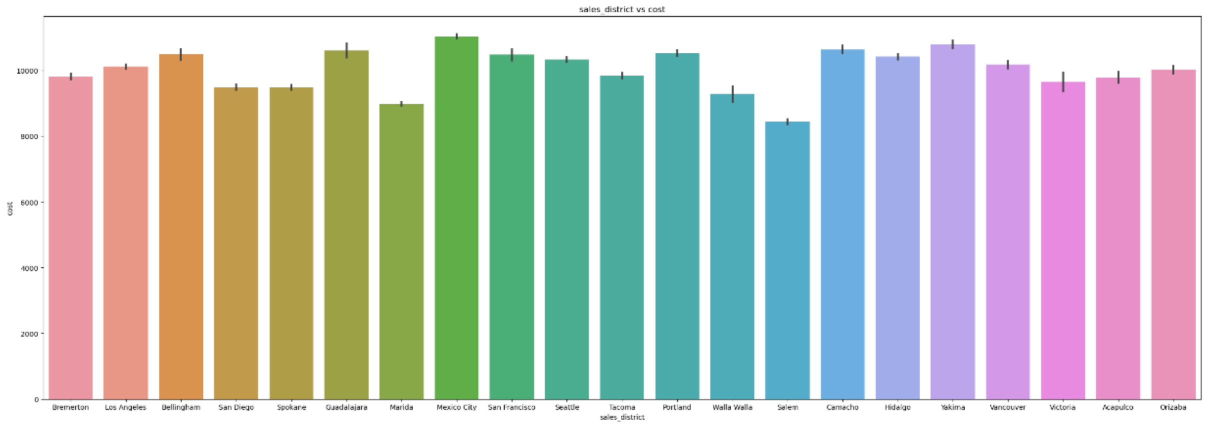
store_cost
unit_sales
SRP
net_weight
cost
store_sqft
promotion_period
1
0.48
0.79
0.03
-0.0016
0.017
-0.0056
1
-0.0015
0.002
-0.01
0.029
0.0017
0.037
0.0028
0.0014
-0.0081
0.00016
-0.004
0.00047
-0.037
-0.12
-0.015
1.0
0.8
0.6
0.4
0.2
0.0

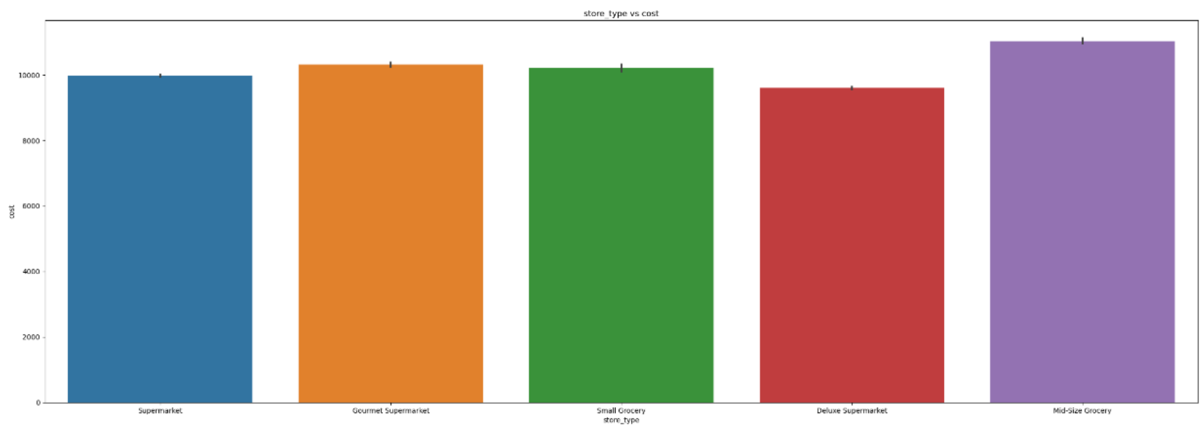

cost vs promotion_period
cost
promotion_period
14000
12000
10000
8000
6000
0
5
10
15
20
25
30

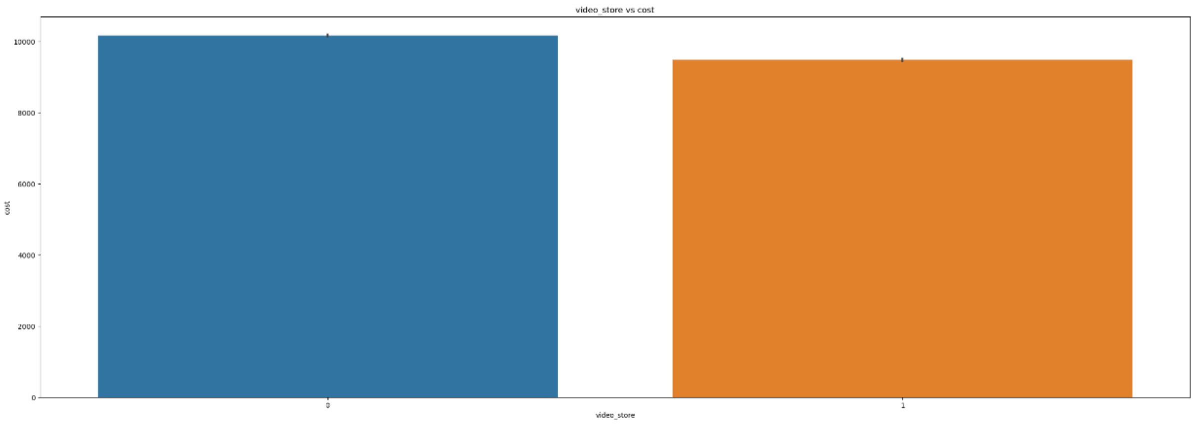

cost vs store_sqft
cost
store_sqft
14000
12000
10000
8000
6000
20000
22500
25000
27500
30000
32500
35000
37500
40000


media_type vs cost
0
2000
4000
6000
8000
10000
Cash Register Handout
Bulk Mail
Daily Paper
Street Handout
Daily Paper, Radio
Product Attachment
Sunday Paper, Radio
Radio
In-Store Coupon
Sunday Paper
TV
Daily Paper, Radio, TV
Sunday Paper, Radio, TV
media_type

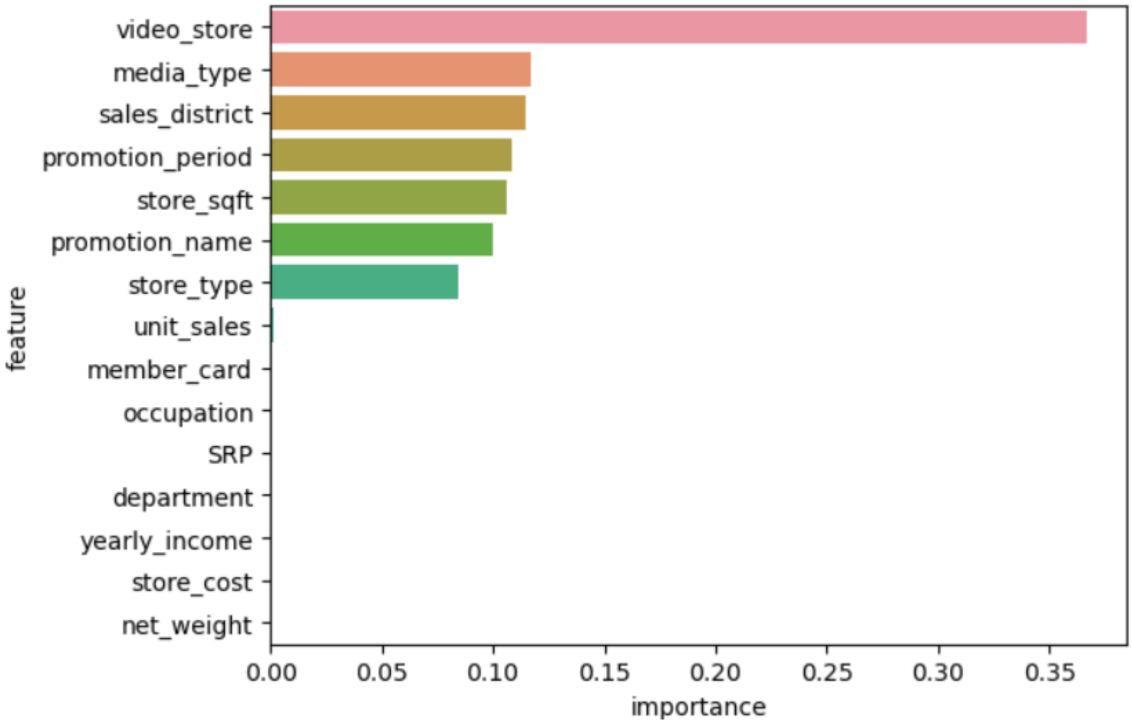

promotion_name vs cost
0
2000
4000
6000
8000
10000
12000
14000
cost
promotion_name


sales_district vs cost
0
2000
4000
6000
8000
10000
cost
Bremerton
Los Angeles
Bellingham
San Diego
Spokane
Guadalajara
Merida
Mexico City
San Francisco
Seattle
Tacoma
Portland
Walla Walla
Salem
Camacho
Hidalgo
Yakima
Vancouver
Victoria
Acapulco
Orizaba
sales_district

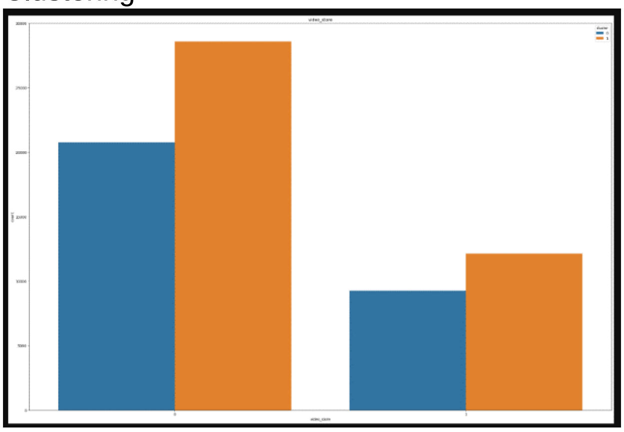

store_type vs cost
0
2000
4000
6000
8000
10000
cost
Supermarket
Gourmet Supermarket
Small Grocery
Deluxe Supermarket
Mid-Size Grocery
store_type

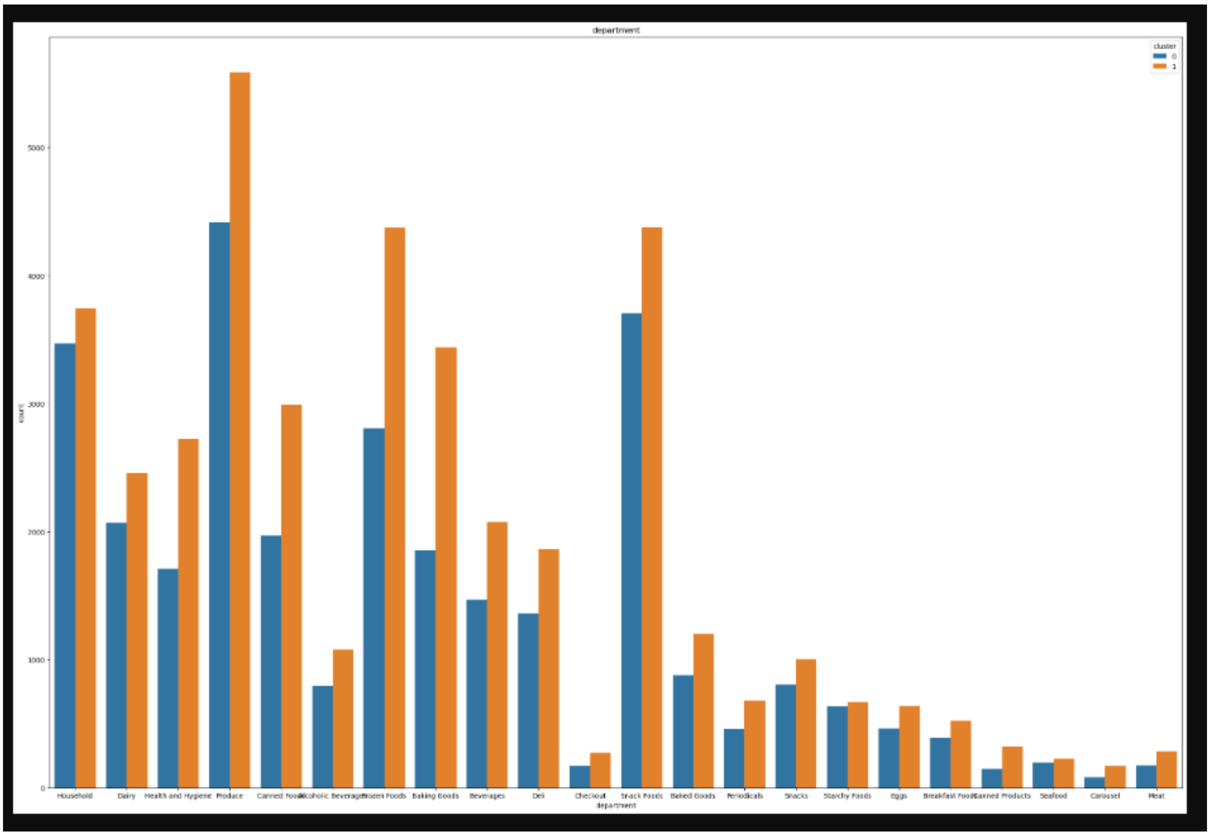


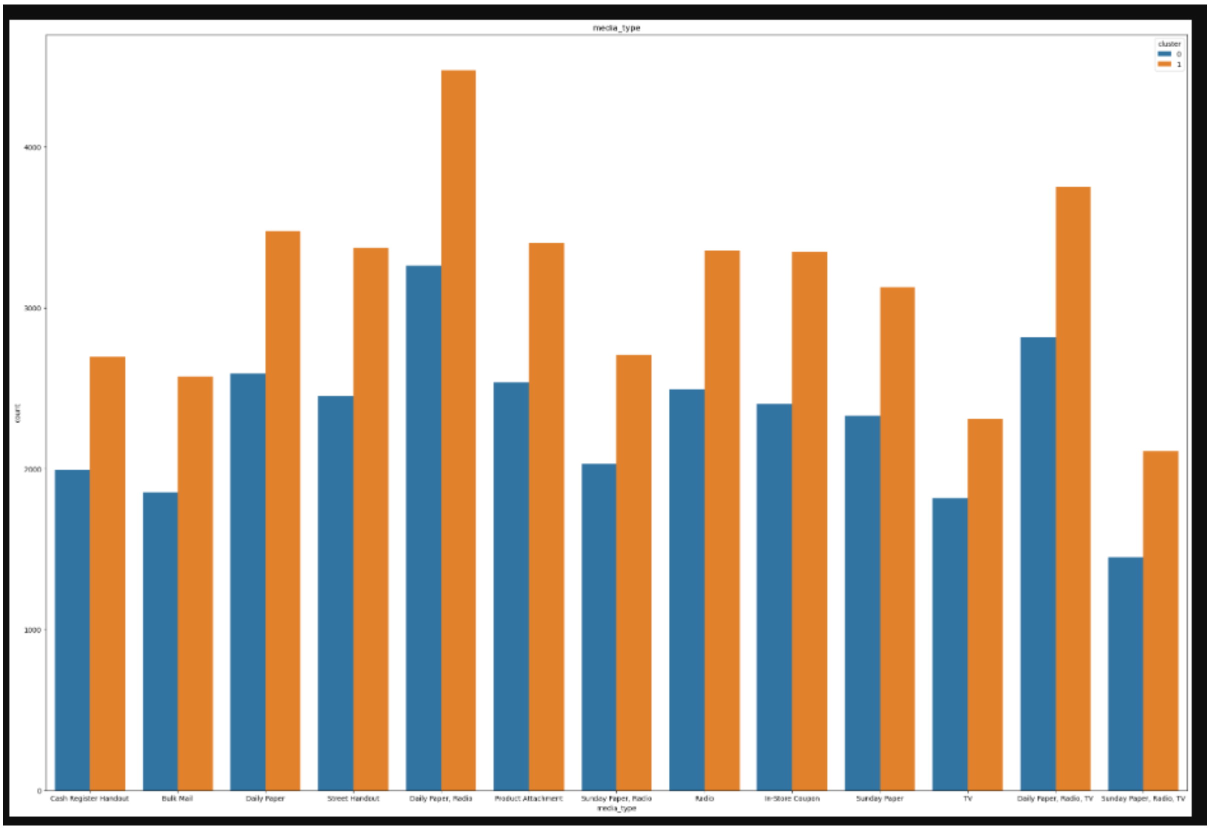
b. Feature Importance of Final Model

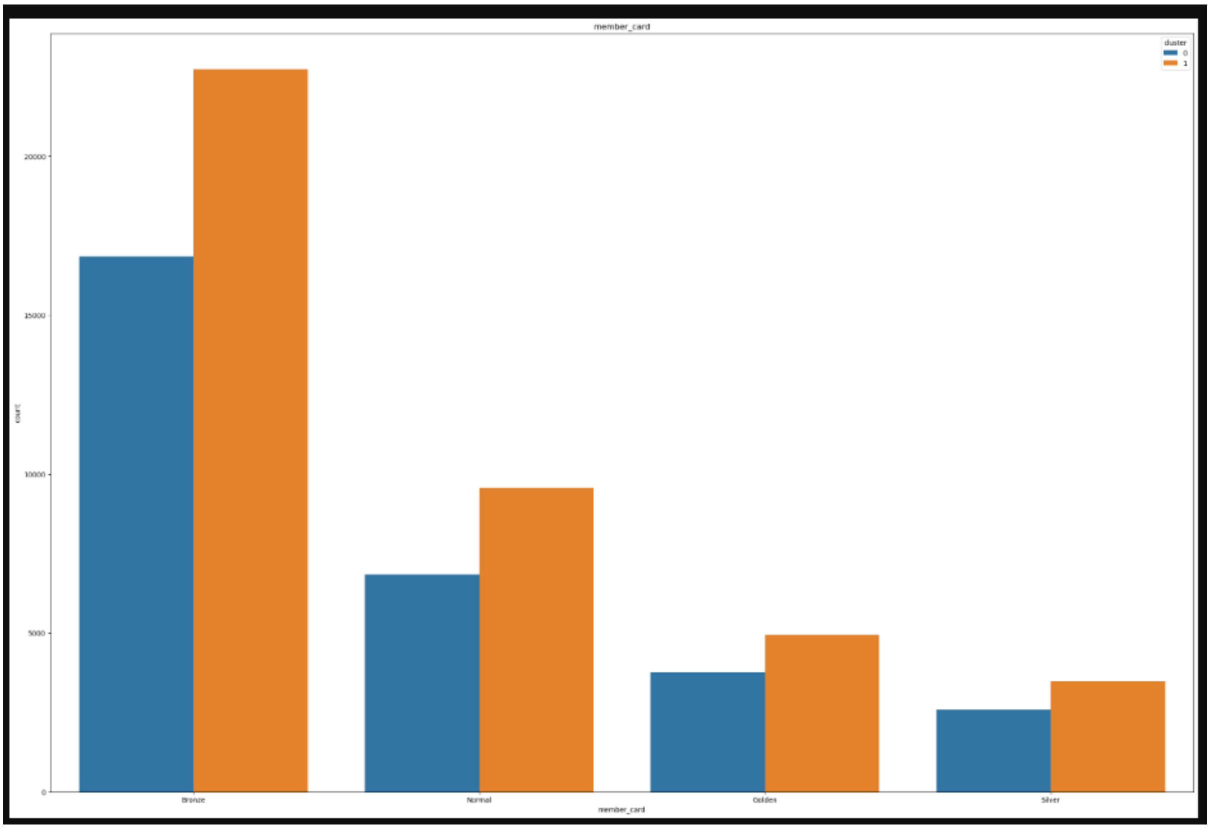


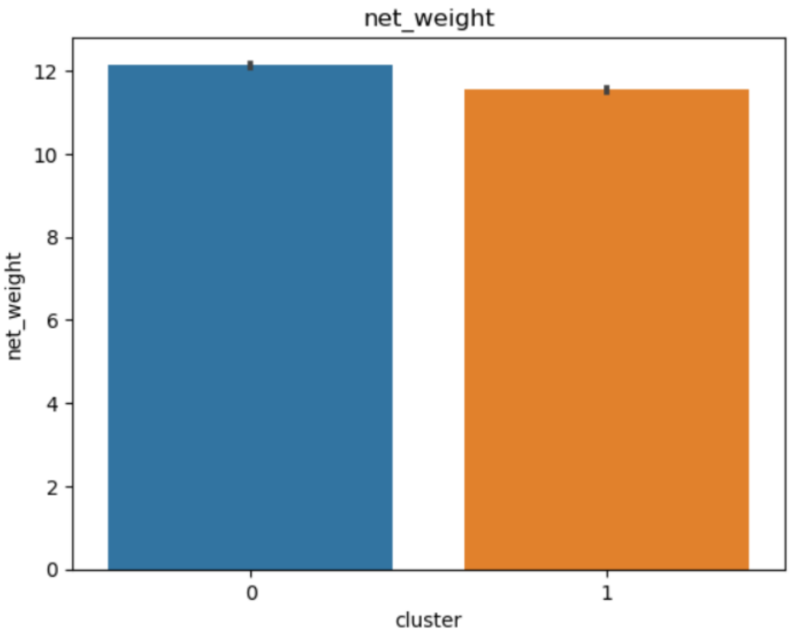
2. Clustering

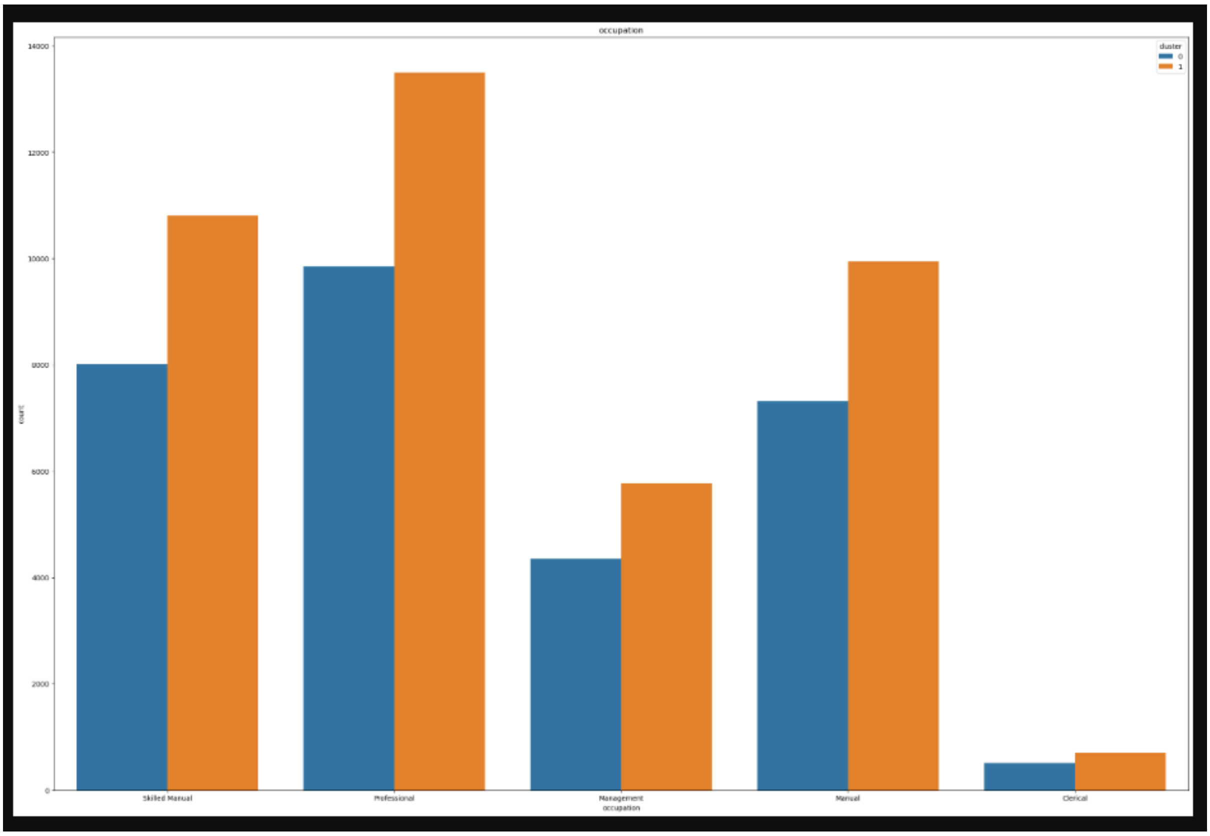


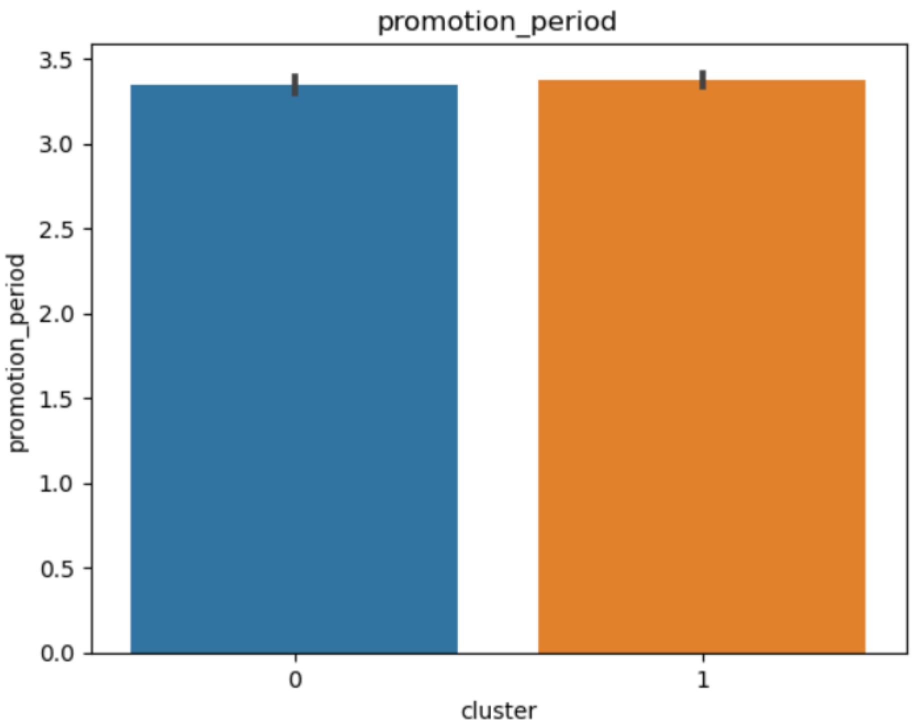

department
cluster
0
1
Household
Dairy
Health and Hygiene
Produce
Canned Foods
Alcoholic Beverages
Frozen Foods
Baking Goods
Beverages
Deli
Checkout
Snack Foods
Baked Goods
Periodicals
Snacks
Starchy Foods
Eggs
Breakfast Foods
Canned Products
Seafood
Carousel
Meat
department

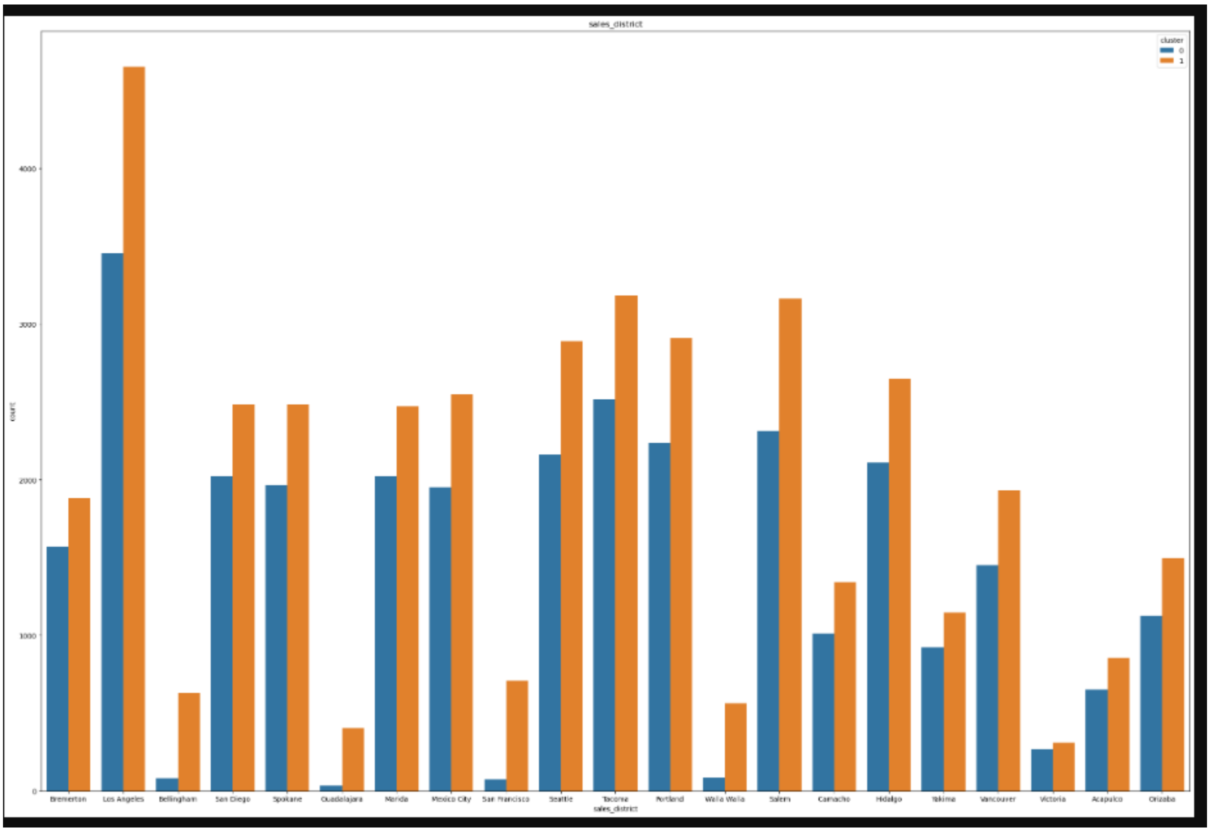

media_type
cluster
0
1
Cash Register Handout
Bulk Mail
Daily Paper
Street Handout
Daily Paper, Radio
Product Attachment
Sunday Paper, Radio
Radio
In-Store Coupon
Sunday Paper
TV
Daily Paper, Radio, TV
Sunday Paper, Radio, TV
media_type

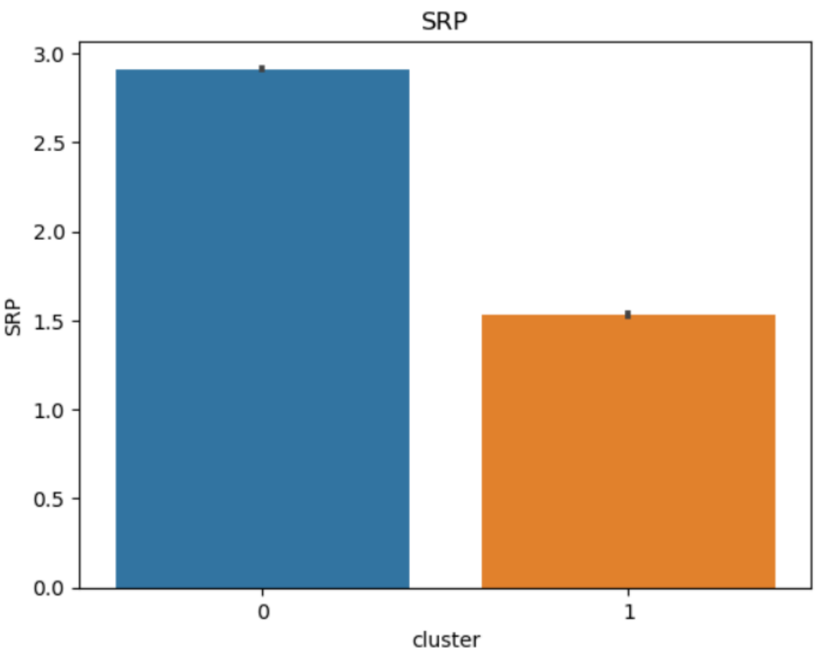

member_card
cluster
0
1
20000
15000
10000
5000
0
count
Bronze
Normal
Golden
Silver
member_card

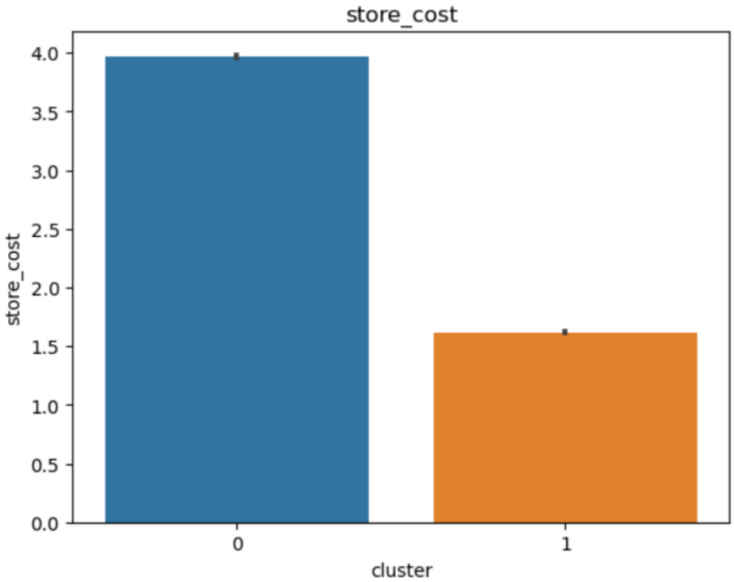

net_weight
12
10
8
6
4
2
0
net_weight
0
1
cluster

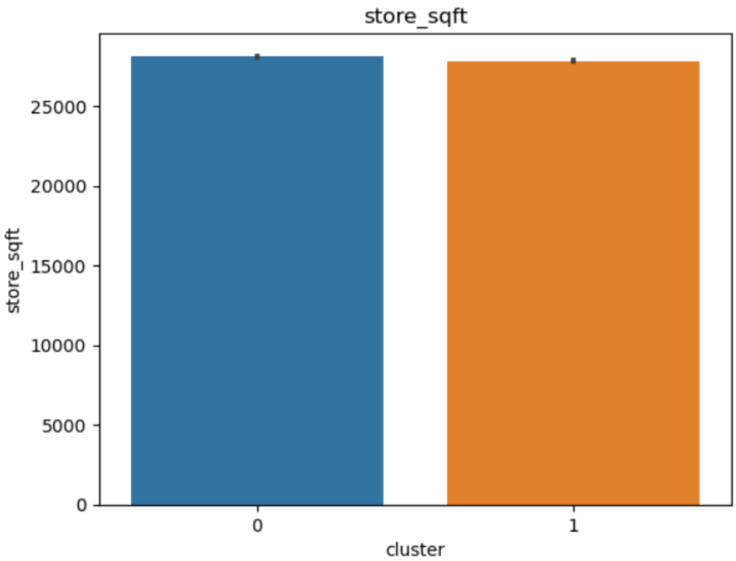

occupation
cluster
0
1
14000
12000
10000
8000
6000
4000
2000
0
count
Skilled Manual
Professional
Management
Manual
Clerical
occupation

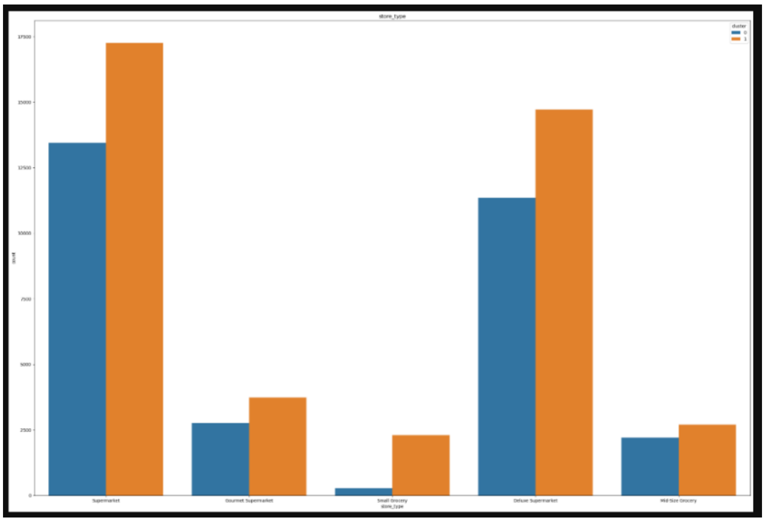

promotion_period
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
promotion_period
0
1
cluster

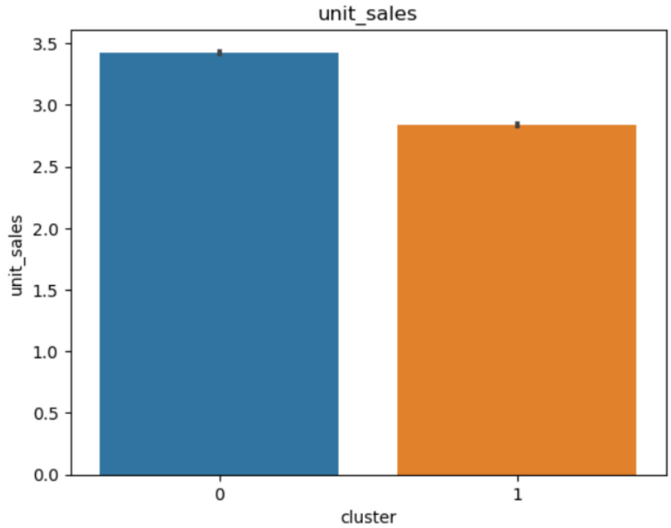

sales_district
cluster
0
1
count
0
1000
2000
3000
4000
Bremerton
Los Angeles
Bellingham
San Diego
Spokane
Guadalajara
Merida
Mexico City
San Francisco
Seattle
Tacoma
Portland
Walla Walla
Salem
Camacho
Hidalgo
Yakima
Vancouver
Victoria
Acapulco
Orizaba
sales_district

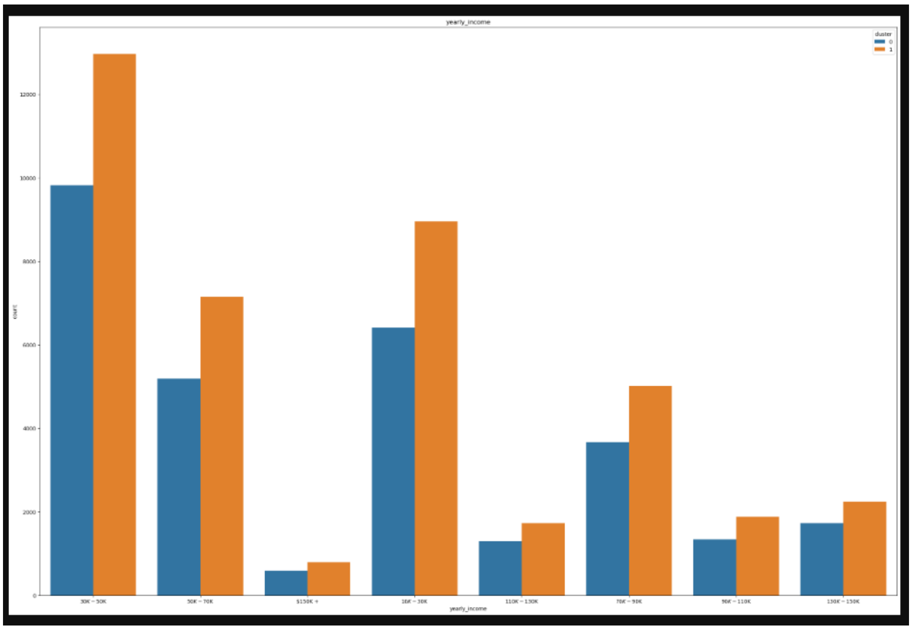

SRP
SRP
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
cluster


store_cost
store_cost
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
cluster


store_sqft
store_sqft
25000
20000
15000
10000
5000
0
0
1
cluster







Implications

1. Commercial Value
   Accurate prediction of customer acquisition cost can help in effecive marketing spend and in cash crunch situation, could help streamlining/prioritizing customers based on the cost of acquisition. Similarly, we can move from CAC to CPC(Cost per acquisition) by analysis CAC for different marketing mediums for prioritizing marketing channels.
   This project while for a chain of brick-and-mortar stores could be extrapolated to companies operating in ecommerce space (by adding ecommerce and online methods and metrics).
2. Recommendation

a. Most important features that are important in predicting customer acquisition cost are video store (presence or absence), media type, sales district, promotion period, store area (square feet), promotion name and store type.
b. Smaller customer acquisition cost for smaller store area.
c. Reduce promotion period to reduce customer acquisition cost
d. Stores having video stores to reduce customer acquisition cost
e. Promotion channel (media type) of Daily Paper, Radio and TV together with respect to Bulk Mail can be used to reduce customer acquisition cost
f. Promotion method of Cash Register Lottery with respect to Bag Stuffers (discounts on items bought in bulk) can be used to reduce customer acquisition cost
g. Allott less customer acquisition cost in stores in Bremerton (Washington) with respect to Acapulco (Mexico)
h. Segmentation of customers into 2 groups can help in subsequent targeting and positioning via different media channels would result in efficient allocation of customer acquisition cost:

   1. price discount searchers buying costlier products and more number of products.
   2. sale day searchers buying cheaper products and less number of products.

Limitations

1. Train and Test MSE are not near zero (51.58 and 56.16 respectively)
2. The clusters or segments formed are not clearly distinct. Some features are common amongst segments while some features are different. This might result in extra spending of resources while making segment specific plans for acquiring customers.
3. Promotion channels used in our dataset and hence in our model are from the era of late 90s. Today's promotion channels are mostly online and hence are easily trackable and thus make marketing promotion spend for acquiring customers more efficient which are not included in our dataset and model.

a. For enhancing our solution, we could update our model to include more current promotion channels (include digital) for a solution which is more up to date in real world.

Closing Reflections

a. Learnings

   1. Optimal feature selection can provide ideal complexity which provides minimum bias error and variance (data sensitivity) error which results in effective and generalized model.
   2. K-prototype clustering used for mixed data uses distances for scaled numerical data (cluster centroid) and matching dissimilarity(no of dissimilarities for all features in each observation from cluster mode) measure for categorical data(cluster mode).

greatlearning

b. Improvements for next iteration

1. Usage of Hyperparameter tuning for the final model for improving variance error and possibly bias errors (no parameters were used for all models).
2. Usage of K-Means clustering for numerical features and K-Modes clustering for categorical features and thereafter comparison with K-Prototype clustering done on both numerical and categorical features.

References:

https://www.shopify.com/blog/customer-acquisition-cost

https://www.zendesk.com/in/blog/customer-acquisition-cost/

https://www.geckoboard.com/best-practice/kpi-examples/customer-acquisition-cost/

https://www.geckoboard.com/best-practice/kpi-examples/ltv-cac-ratio/

https://www.geckoboard.com/best-practice/kpi-examples/average-order-value/

https://www.geckoboard.com/best-practice/kpi-examples/percent-returning-customers/

https://blog.converted.in/en-us/blog/average-customer-acquisition-cost-in-ecommerce

https://www.linkedin.com/pulse/true-cost-customer-acquisition-within-fmcg-how-lower-/