AI for Business Decisons Final Project

December 2024

Using dunnhumby data to scale MAYBELLINE NEW YORK

Ananth Mohan
Chhaya Tundwal

Understanding how Maybelline can improve within the make-up market

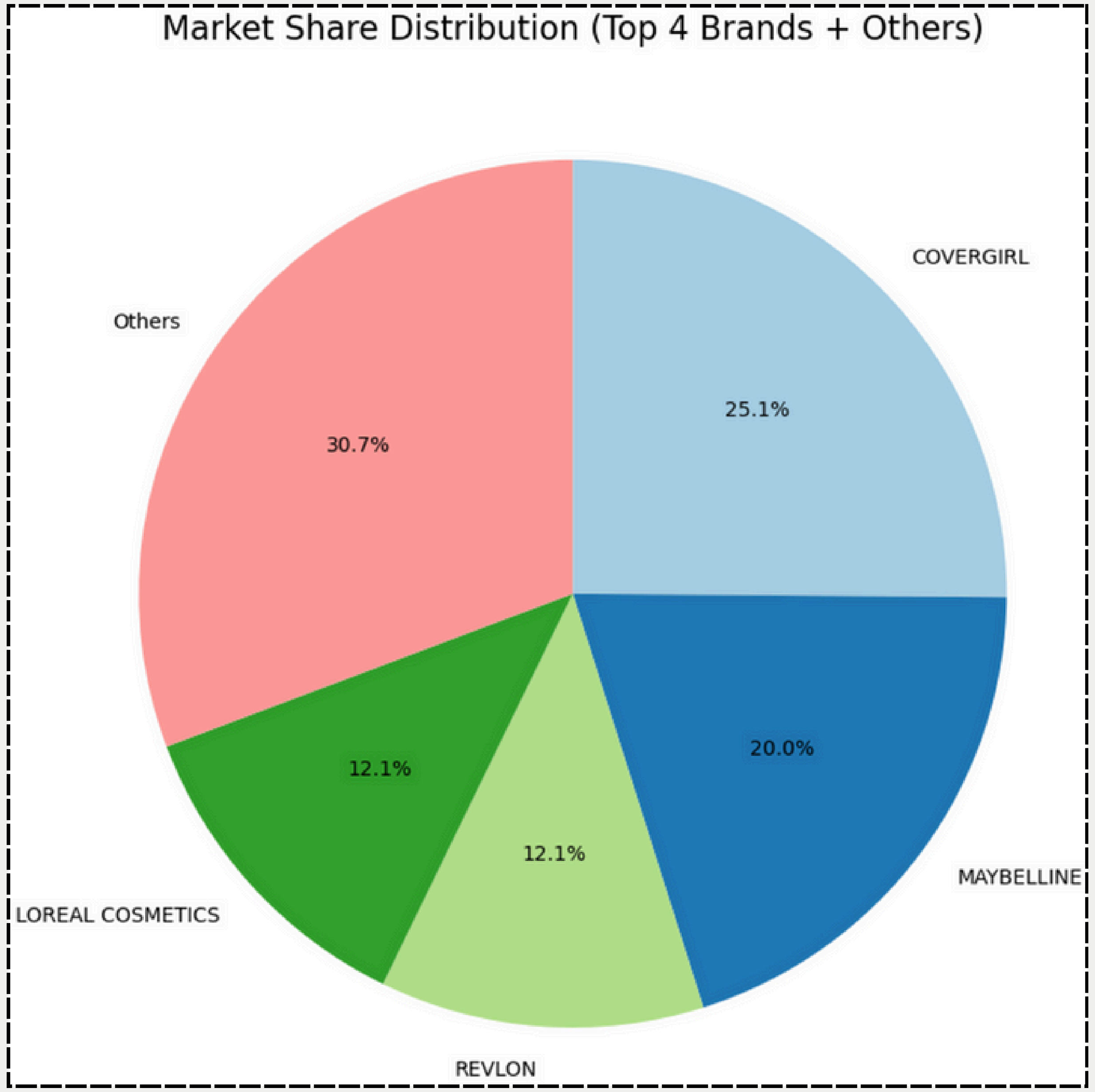
Today's Agenda

1. Problem Statement
2. Approach
3. Models
4. Summary and Recommendations

Background

Maybelline has a market share of ~20%*
in a moderately concentrate market
(HHI Index = 0.15)

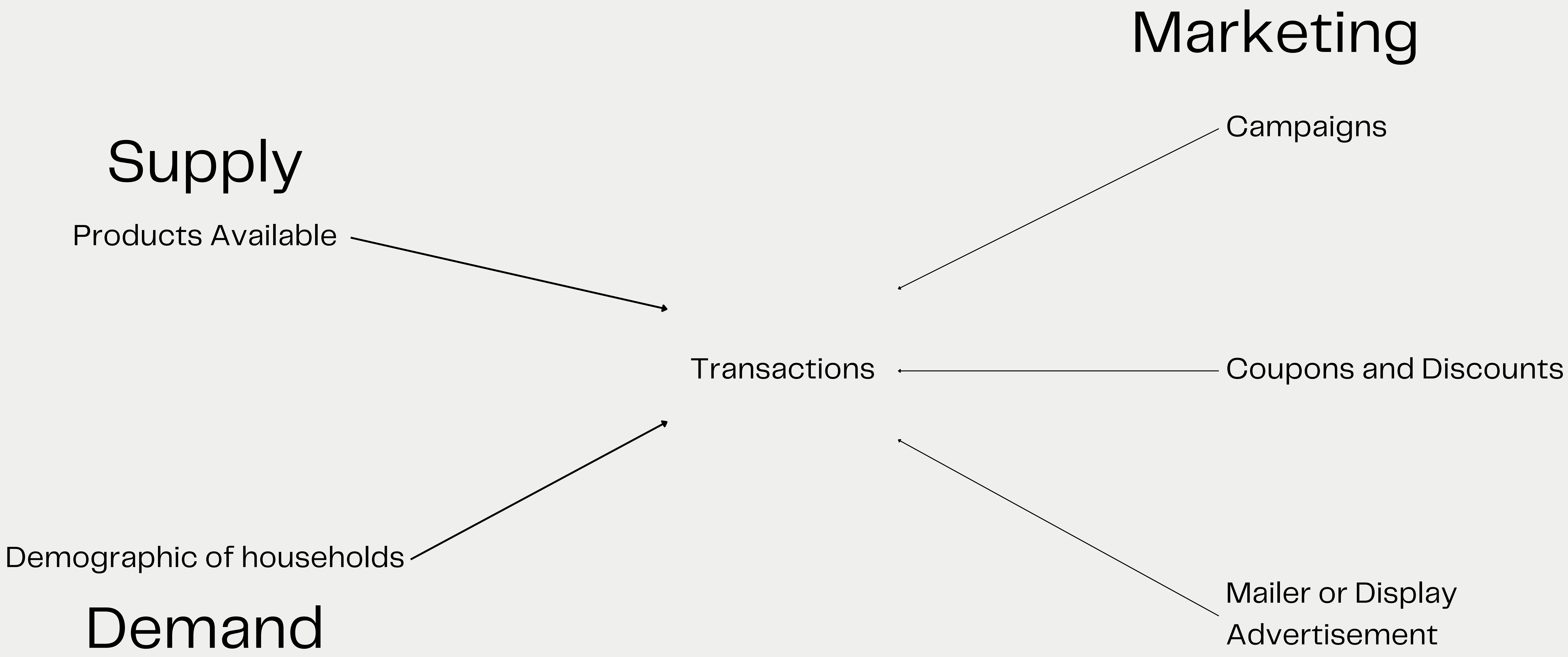
Maybelline is a multinational company focusing on cosmetics, skin care, perfume, and personal care



So why is Maybelline at 20% and can they go up?

* Based on Dunnhumby data for 2 years; HHI – Herfindal Hirschman Index

How will Dunnhumby data help us



How Dunnhumby Empowers Us – and Its Limitations

1. Comprehensive Transaction Data, for 2 years
2. Marketing Channels and their data
3. Summarised Demographics data

1. Imbalanced Data
2. Limited variables. Real world influencers are not captured
3. Lack of interpretability with some of the demographic variables
4. Single Retail Chain

General Approach

EDA

- Class Imbalances
- Data types, Missing Values, Distributions

Preprocessing & Transforming

- Outlier Processing
- Feature Transformation
- Removing correlated columns

Model

- Linear Regression
- Logit Regression
- Gradient Boosted Trees
- Random Forest
- K-Means Clustering

Evaluation & Feature Selection

- Based on R-square and MAPE values

Insights

- Converting these numbers to business actions

Our Solutions

Retrospective	Marketing Channels	Which among the 3 marketing channels are most effective – Display ads, Weekly Mailers, and Coupons through Campaign
	Who are the customers?	What are main customer/ household personas who buy Makeup products?
Predictive	Demand Forecasting	How to forecast demand for makeup products?
	Product Recommendation	Can Maybelline increase it’s revenue by bundling products together?

Problem #1

Marketing Channels

Use regression to understand the influence of marketing efforts on sales

OLS Regression Results

Dep. Variable:	transaction_count	R-squared:	0.005
Model:	OLS	Adj. R-squared:	-0.006
Method:	Least Squares	F-statistic:	0.4407
Date:	Sun, 08 Dec 2024	Prob (F-statistic):	0.927
Time:	19:15:19	Log-Likelihood:	-3.9934
No. Observations:	920	AIC:	29.99
Df Residuals:	909	BIC:	83.05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.9257	0.108	8.579	0.000	0.714	1.138
mailer[T.A]	0.1251	0.108	1.161	0.246	-0.086	0.336
mailer[T.C]	0.1243	0.121	1.027	0.305	-0.113	0.362
mailer[T.H]	0.1198	0.115	1.038	0.299	-0.107	0.346
display[T.1]	0.0743	0.204	0.364	0.716	-0.326	0.474
display[T.3]	0.0564	0.131	0.431	0.666	-0.200	0.313
display[T.4]	0.0743	0.154	0.483	0.629	-0.227	0.376
display[T.5]	0.0222	0.113	0.196	0.845	-0.200	0.244
display[T.6]	0.0743	0.154	0.483	0.629	-0.227	0.376
display[T.7]	0.1664	0.112	1.489	0.137	-0.053	0.386
display[T.9]	0.1992	0.131	1.523	0.128	-0.057	0.456

OLS Regression Results

Dep. Variable:	total_quantity	R-squared:	0.648
Model:	OLS	Adj. R-squared:	0.647
Method:	Least Squares	F-statistic:	1607.
Date:	Sun, 08 Dec 2024	Prob (F-statistic):	0.00
Time:	19:00:56	Log-Likelihood:	-3826.8
No. Observations:	1750	AIC:	7660.
Df Residuals:	1747	BIC:	7676.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.5480	0.056	27.650	0.000	1.438	1.658
retail_disc	-0.2764	0.007	-41.841	0.000	-0.289	-0.263
coupon_disc	-0.3680	0.035	-10.649	0.000	-0.436	-0.300

Omnibus:	1626.312	Durbin-Watson:	1.773
Prob(Omnibus):	0.000	Jarque-Bera (JB):	74740.659
Skew:	4.310	Prob(JB):	0.00
Kurtosis:	33.834	Cond. No.	10.8

Mailer and Display has low significance, but we found a higher influence of discounts on the sales quantity

The **logistic regression model** predicts whether a household will respond to a campaign or not respond.

Target Variable (responded):
1: The household responded to the campaign.
0: The household did not respond to the campaign.

Features:
"Total sales" : How much a household has spent overall.
"Total coupons used": The sum of coupon discounts redeemed by the household.
"Average discounts": The average retail discount availed.
"Campaign duration": The length of the campaign.

Overall Performance: The model has high accuracy, a solid ROC-AUC score of 0.95, and performs quite well on both precision and recall, making it a strong model for predicting customer response.

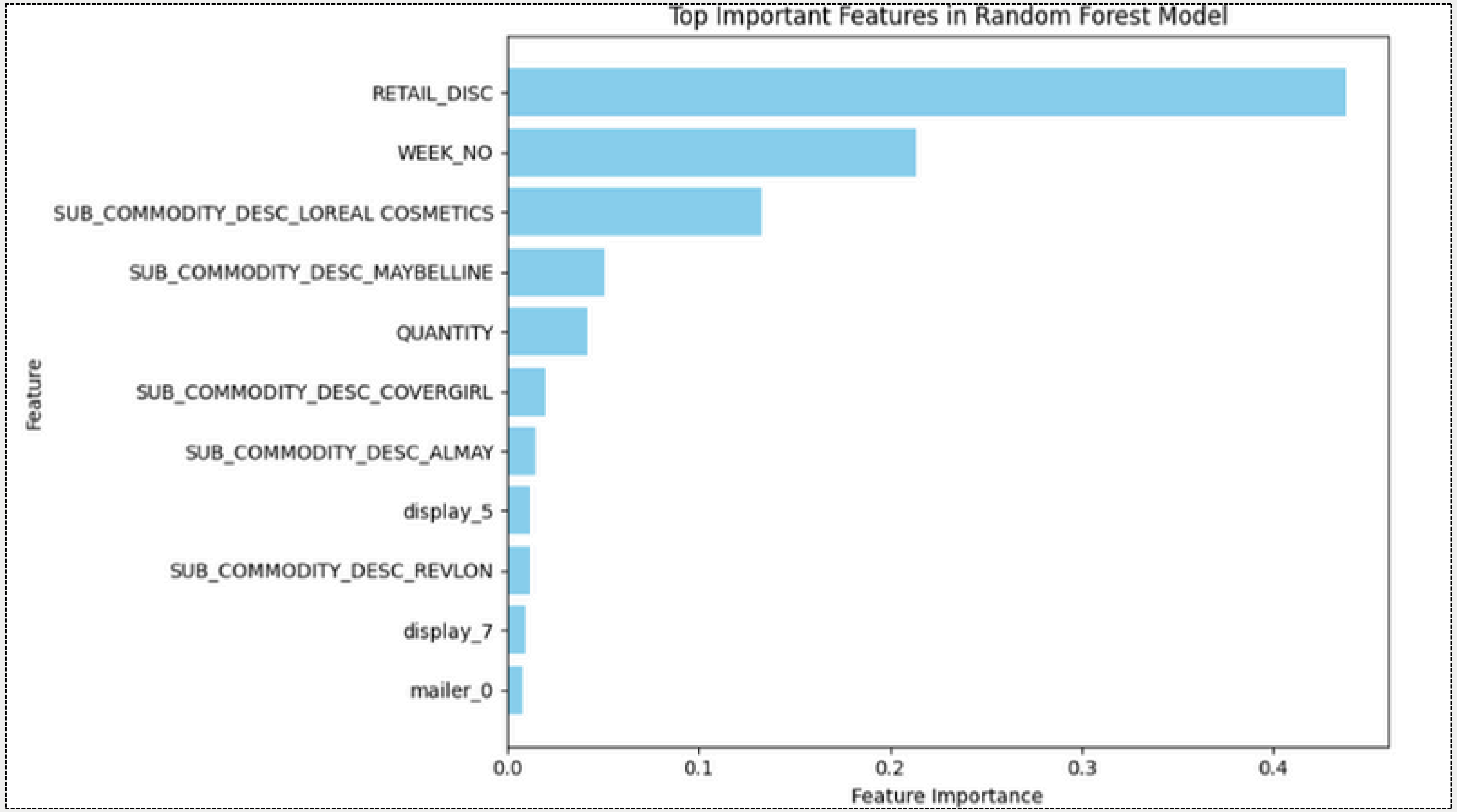
95%

ROC-AUC

	precision	recall	f1-score
0	0.91	1.00	0.95
1	1.00	0.87	0.93
accuracy			0.94
macro avg	0.96	0.93	0.94
weighted avg	0.95	0.94	0.94

ROC-AUC Score: 0.95

Random Forest model to predict the sales for products that belong to "COSMETICS" category.



Retail Discounts and Timing are Dominant: The model highlights that offering retail discounts and understanding the week of the transaction are critical drivers.

Brand-Level Influence: indicating competitive dynamics and brand-specific promotions significantly impact outcomes.

Quantity's Moderate Role: Bulk purchases have some influence, reflecting customer purchasing patterns.

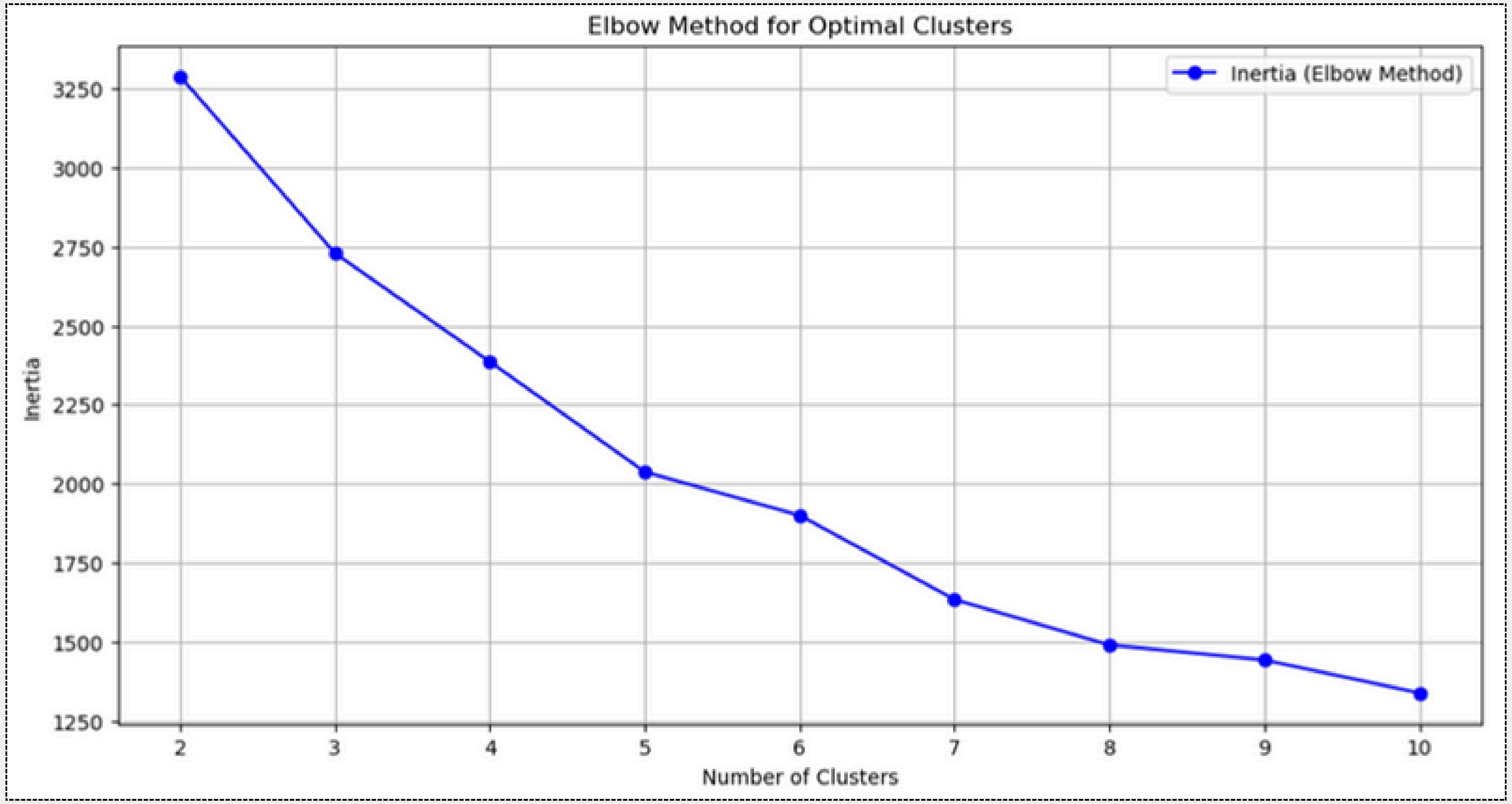
Opportunity to Improve Mailer/Display Effectiveness: The lower importance of mailer and display features suggests these strategies may not be optimized or have a smaller impact on customer decisions.

Problem #2

Who are the customers?

Using cluster analysis (K-Means Clustering) to identify the group of customers based on demographics and their spending habits

Used Elbow method to identify the recommended number of customer segments



From the graph, we can infer that k=3, 4, or 5 could be optimal cluster values. After further experimentation, we determined that k=4 is the most suitable choice for our category.

Cluster 0: High-value customers, frequent buyers, highly engaged with campaigns.

Cluster 1: Low-spending, disengaged customers. Potential opportunity for acquisition or reactivation.

Cluster 2: Moderately engaged customers who may respond to discounts and campaigns.

Cluster 3: Disengaged customers with minimal spending; potential focus group for re-engagement.

Cluster	HOMEOWNER_DESC	KID_CATEGORY_DESC	total_spend	avg_spend	purchase_count	total_retail_disc	avg_retail_disc	total_coupon_disc	avg_coupon_disc	campaigns_engaged
0	Homeowner	None/Unknown	30.077407	1.075021	27.888889	-8.902222	-0.320917	0.000000	0.000000	7.518519
1	Unknown	None/Unknown	2.851062	0.758532	2.522124	-0.809115	-0.223405	0.000000	0.000000	4.862832
2	Homeowner	None/Unknown	6.824486	1.333861	5.588785	-2.122523	-0.443716	-0.017477	-0.002999	5.789720
3	Homeowner	None/Unknown	1.110778	0.414996	1.110778	-0.303353	-0.112514	0.000000	0.000000	5.005988

Recommendation : The output suggests prioritizing Cluster 0 and 2 for retention and growth strategies while considering campaigns to re-engage Cluster 1 and 3.

Problem #3

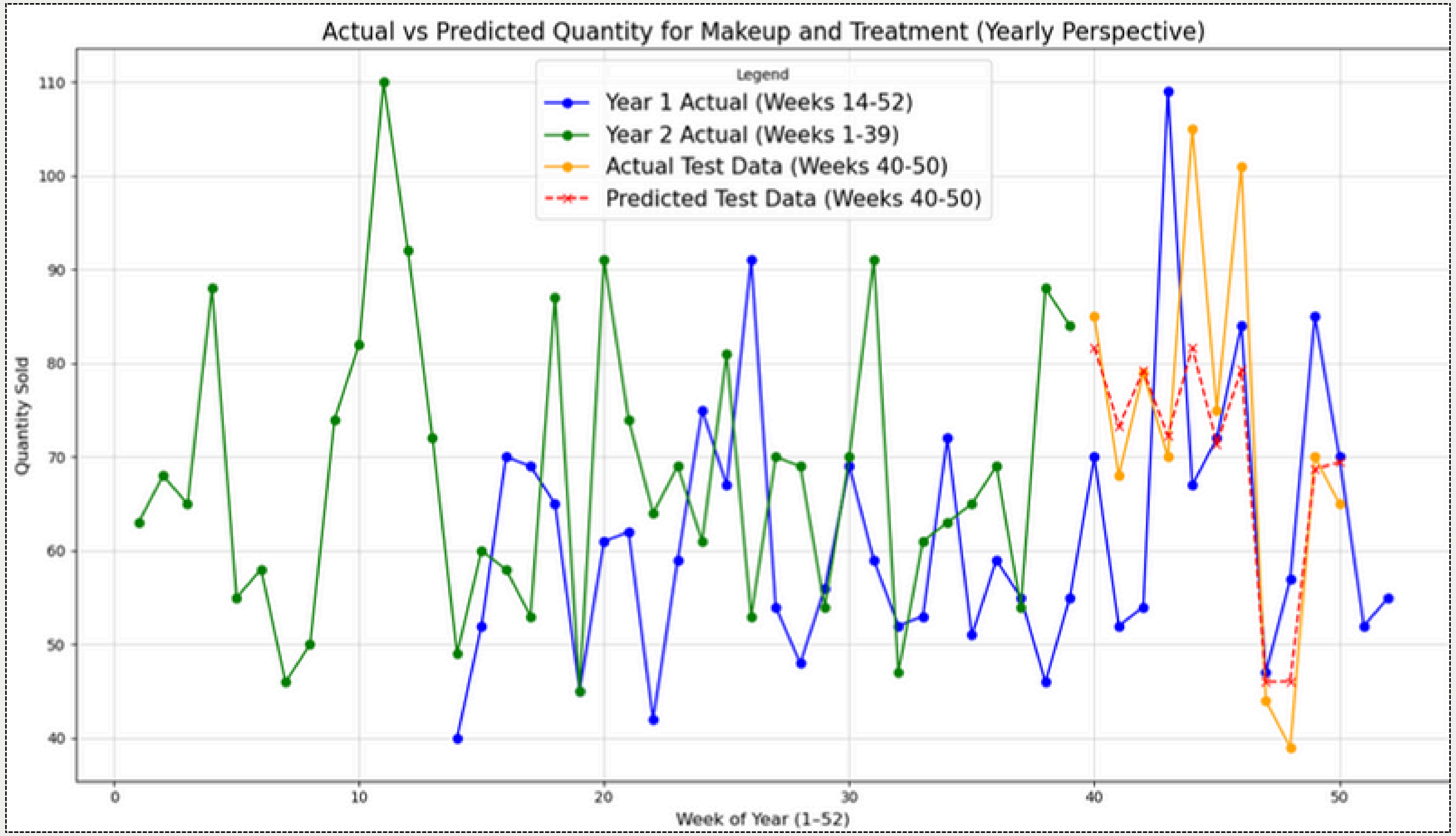
6%

Training MAPE

8.6%

Test MAPE

Use a gradient boosting model to predict last 3 months sales based on on first 91 weeks of data



Demand Forecasting

Features Used:

- ‘1 week lagger’
- ‘4 week lagger’
- ‘8 week lagger’
- 'month'
- 'year'
- 'sales_value'
- 'retail_disc'
- 'coupon_disc'

Recommendation : Use the model to predict demand, and stock appropriately

Problem #4

Market Basket analysis to identify and analyze product pairs frequently purchased with Maybelline products

	Product_Pair	Count
19692	(1029743, 1082185)	107
8134	(981760, 1082185)	82
15264	(1082185, 1127831)	81
70806	(951590, 1082185)	75
19110	(866211, 1082185)	71
2517	(862349, 1082185)	68
23659	(1082185, 1126899)	66
663	(995242, 1082185)	66
3759	(961554, 1082185)	66
989	(1070820, 1082185)	62

Product Recommendation

A higher percentage indicates a greater likelihood that buyers will perceive value in this bundle

	Maybelline_Product	Most_Frequent_Product	Percentage
0	9796730	1137808	2.083333
1	1060119	824072	7.692308
2	915800	830795	25.000000
3	923552	1101706	2.000000
4	6396131	849330	4.166667
..	...	...	...
497	9530255	34873	25.000000
498	1102188	866540	16.666667
499	10456573	840890	12.500000
500	10457517	840890	12.500000

Recommendation : Products with a higher likelihood of being purchased together can be promoted using coupons to incentivize buyers

Executive Summary

Objective:

This project aims to understand what drivers affect sales of makeup products in this retailer and how maybelline can improve market share

Output:

Our analysis and recommendations aims to improve product sales and marketing outcomes as well as be better prepared for upcoming demand

Recommendations:

1. **What Marketing works**: Retail discounting works best, so a store which has a higher number of loyal members can increase overall sales for all products.
2. **Who buys Makeup products**: Focusing more on the 2 customer segments, i.e., the high spending frequent shopper and value conscious coupon spenders can improve sales
3. **What is the seasonality in sales**: Makeup sales are volatile and doesn't follow a strict pattern. Hence using a GBT model can help in managing inventory throughout the year
4. **Product Recommendation**: Bundling together frequently bought products can potentially improve revenue

Thank you!

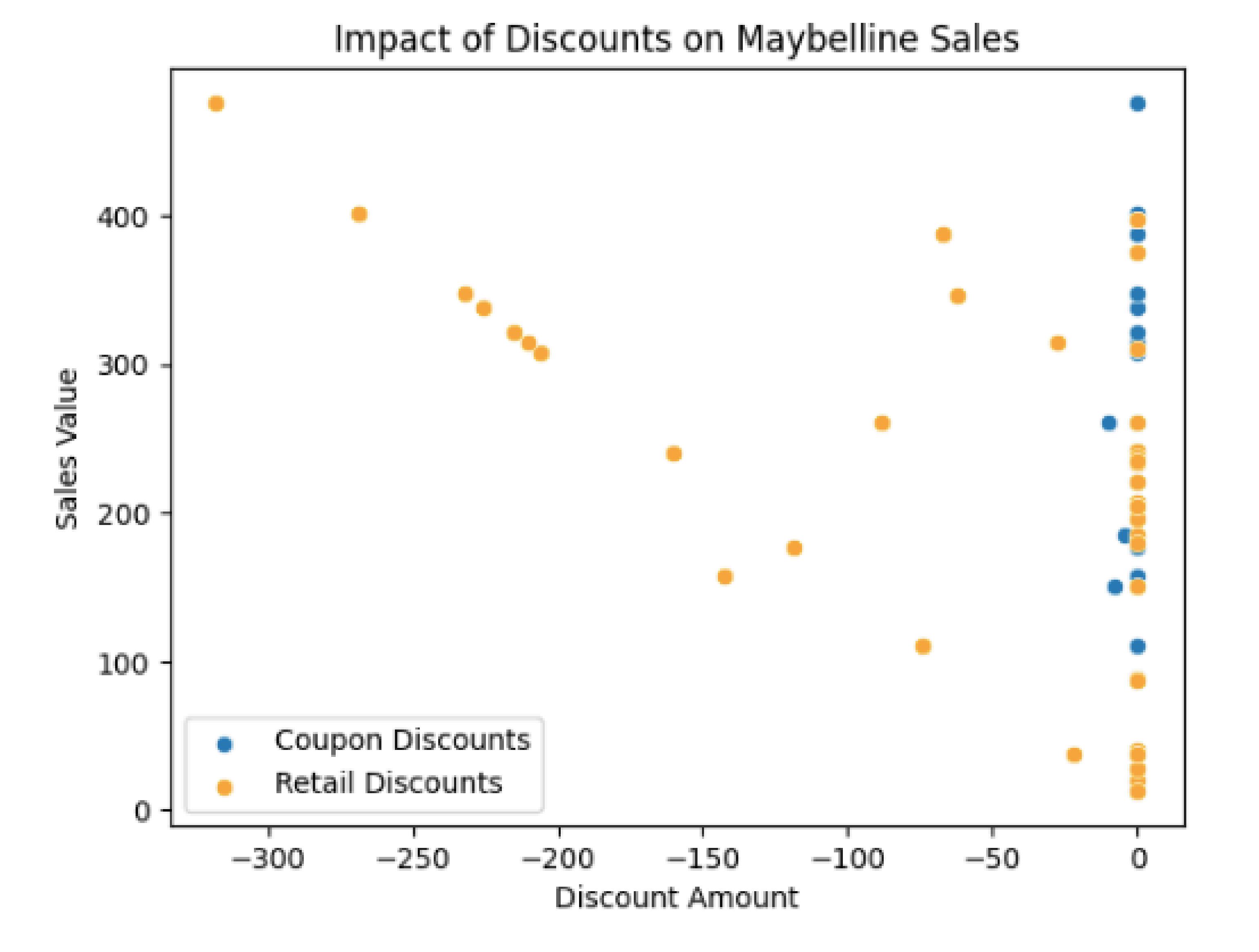
Enjoy your holidays!

Appendix

Comparison of Discounts:
Retail discounts seem to drive higher sales values compared to coupon discounts at most discount levels.

Retail Discounts (Loyalty Programs): These are likely more impactful in driving sales for Maybelline, as higher sales values are concentrated in this category.

Coupon Discounts: Despite offering larger discounts in some cases, coupon discounts seem to generate lower sales values on average.



Recommendation:
Focus on retail loyalty programs, as they seem to have a stronger impact on driving sales.
Consider improving coupon distribution methods or targeting specific customer segments to maximize their usage.

References:

1. Books and Academic Sources:
    - Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers. (For Market Basket Analysis concepts)
    - Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts. (For demand forecasting techniques)
2. Tools and Libraries Used:
    - LightGBM Library for Gradient Boosting: https://lightgbm.readthedocs.io
3. Market Basket Analysis and Association Rules:
    - Agrawal, R., Imieliński, T., & Swami, A. (1993). "Mining Association Rules Between Sets of Items in Large Databases." ACM SIGMOD Conference on Management of Data.
4. Time Series Forecasting:
    - Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). Time Series Analysis: Forecasting and Control. Wiley.
    - Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1998). Forecasting Methods and Applications. Wiley.
5. Online Tutorials and Documentation:
    - LightGBM Demand Forecasting: https://towardsdatascience.com/demand-forecasting-with-lightgbm-2e9612a55c0e
    - Introduction to Market Basket Analysis: https://www.analyticsvidhya.com/blog/2021/10/a-beginners-guide-to-market-basket-analysis/
    - Time Series Analysis with Python: https://machinelearningmastery.com/time-series-forecasting/
6. Other Relevant Research/Case Studies:
    - Use case of Market Basket Analysis in Retail: https://link.springer.com/article/10.1007/s10260-018-0415-5
    - Case studies on promotional analysis and product bundling strategies: https://hbr.org/

Link to actual analysis:

1. https://colab.research.google.com/drive/1g8pEKNTO2ctOuu-EVQgySUgYIOdXxUcj?usp=sharing
2. https://colab.research.google.com/drive/1GVhnaenCLFm7h7MDPmSJdN2eRiJx98Oy?usp=sharing