Marketing Analytics

Determining if high-value customers are likely to respond to our next marketing campaign

Adithya M
Advaith Shankar
Ammar Mustafa
Manasa Maganti
Shashank Rao
Varsha Manju Jayakumar

Agenda

- Defining Our Problem
- Hypothesis Testing
- Defining our High-Level Strategy
- Function RFM Scores
- Model Testing
- Results from Approach 1
- Results from Approach 2
- Results from Approach 3
- Results from K-means Clustering
- Insights

Problem statement: Targeting high-value customers to predict campaign response

Business Goal

To determine if high-value customers will respond to our next marketing campaign within our diverse demographics, variety of goods, enabling more effective resource allocation

Current State

- ➔ Our customer base is diverse, with varying levels of engagement and purchasing behavior, making it challenging to identify high-value customers
- ➔ Our current marketing strategies lack a strong focus on high-value customers, which may lead to inefficient use of resources
- ➔ We don't yet have a structured approach to consistently identify and engage high-value customers

Desired Future State

- ➔ We aim to clearly identify high-value customers by using data-driven insights like spending patterns, purchase frequency, and recency
- ➔ Our marketing efforts will be strategically targeted towards high-value segments to boost engagement, loyalty, and revenue
- ➔ Resources will be allocated more efficiently, with a focus on high-value customers to maximize return on investment

Questions ?

- ➔ Which customer attributes (e.g., purchase frequency, spending patterns) are most predictive of campaign responsiveness?
- ➔ How can RFM scoring enhance our analysis of predicting the customer's response to campaign?

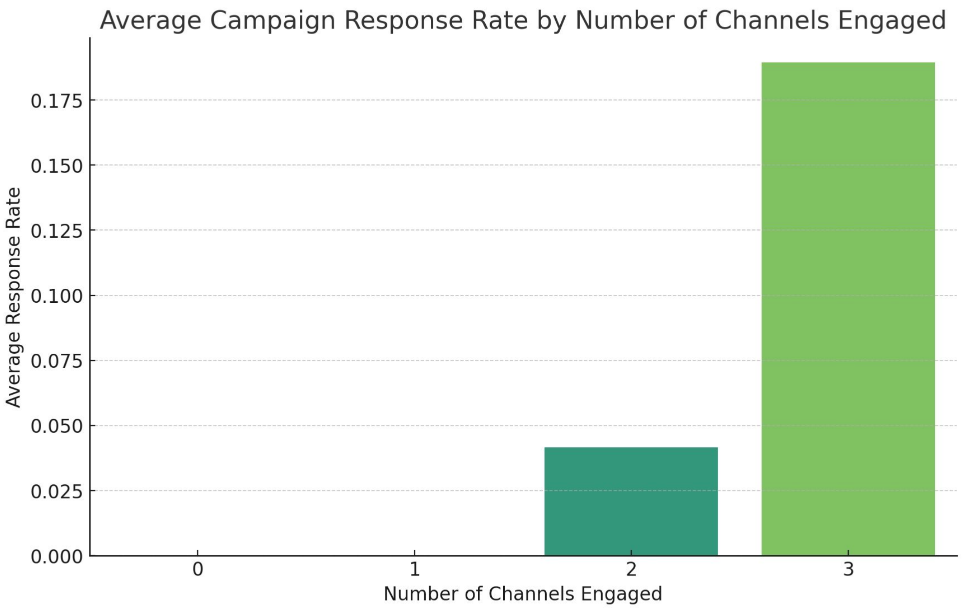
Quick run-through on our Data!

Hypothesis Testing - Time for some fun insights!

Customers who engage through multiple purchase channels are more likely to respond positively to a new campaign

- ➔ Customers using all three channels have the highest response rate (18.9%), showing strong engagement.
- ➔ Those using two channels have a moderate response rate (4.2%).
- ➔ Customers using one or no channels show no recorded responses, indicating low engagement.

Chi-Squared Value: 79.29848852638342
p-value: 4.3401146620042094e-17
Degrees of Freedom: 3



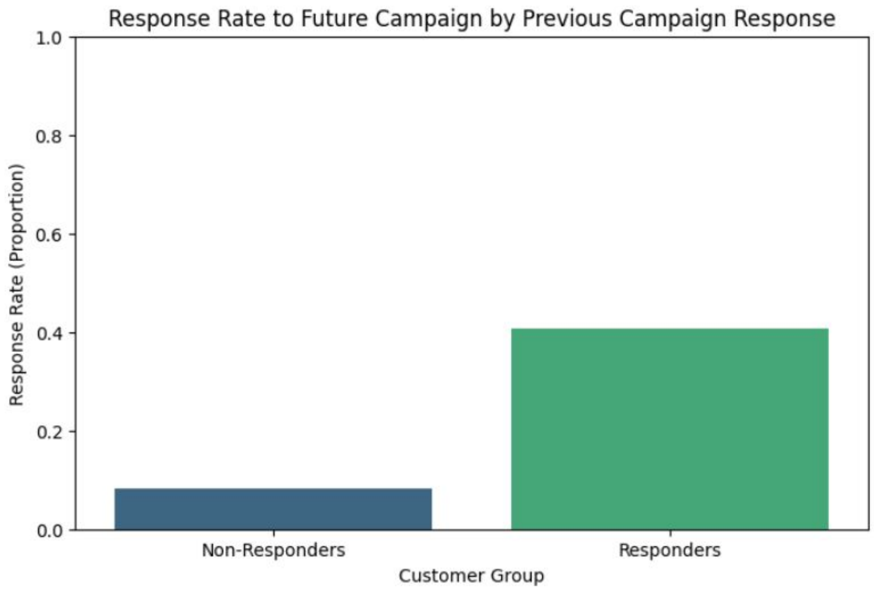
The statistically significant p-value suggests a strong association between engaging through multiple channels and a higher likelihood of responding positively to a marketing campaign, supporting the hypothesis.

Hypothesis Testing - Time for some fun insights!

Customers who engage through previous campaigns are more likely to respond positively to a new campaign

- ➔ Customers who previously engaged with marketing campaigns are more likely to respond positively to future campaigns.
- ➔ This highlights the importance of targeting past responders for future campaigns as they exhibit higher engagement rates.

T-Statistic: 13.632282595948574
P-Value: 1.4280563832017728e-36



The statistically significant p-value suggests customers who responded to previous campaigns are more likely to respond to future campaigns.

Hypothesis Testing - Time for some fun insights!

Customers who are older are more likely to spend more

Old Age Group Spends More:

➔ The **older age group** has a higher average monetary spending and frequency compared to the **young age group**
➔ This suggests that older customers tend to spend more and purchase on products overall

Implications for Marketing:

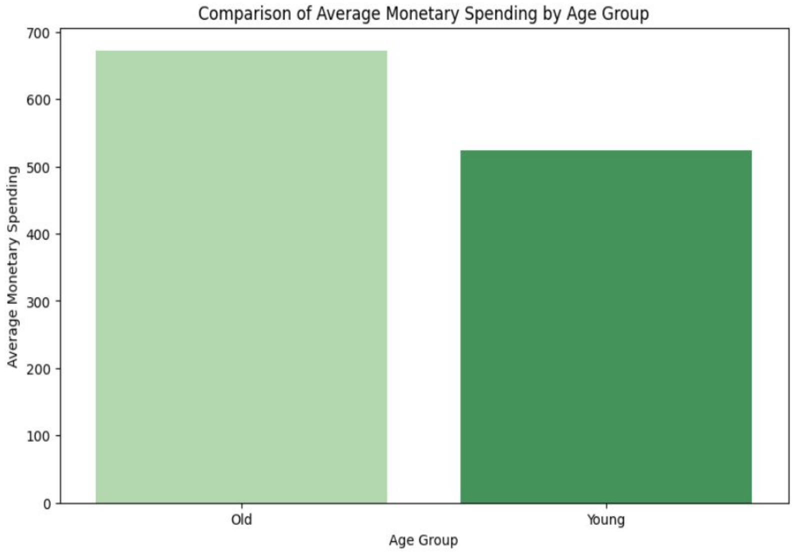
➔ Older individuals may represent a higher-value segment in terms of spending potential.
➔ Marketing strategies could be tailored to target older customers with premium or higher-value product offers.

Frequency

T-statistic: 6.523029299680769 | **P-value:** 1.0749852452000602e-10

Monetary

T-statistic: 5.386460992205868 | **P-value:** 8.957403704315754e-08



The statistically significant p-value to conclude : Older individuals have significantly higher purchase frequency and monetary spending compared to younger individuals.

Hypothesis Testing - Time for some fun insights!

Widows are spending more?

More on what? What about Wine?

➔ Widows have a higher average monetary spending on wine but is it a coincidence?

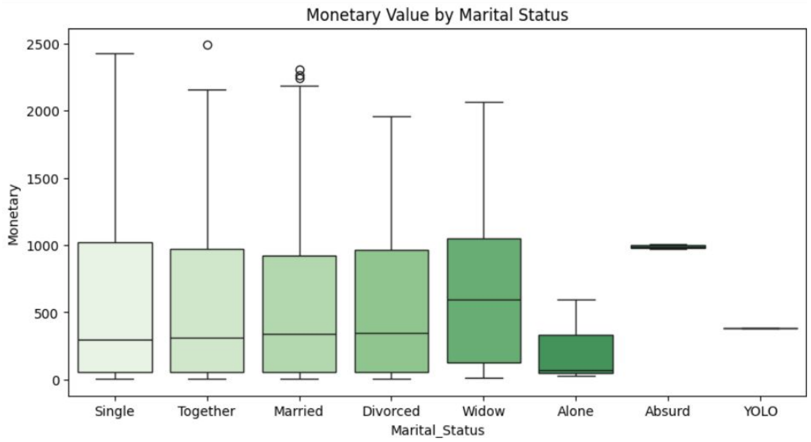
Implications for Marketing:

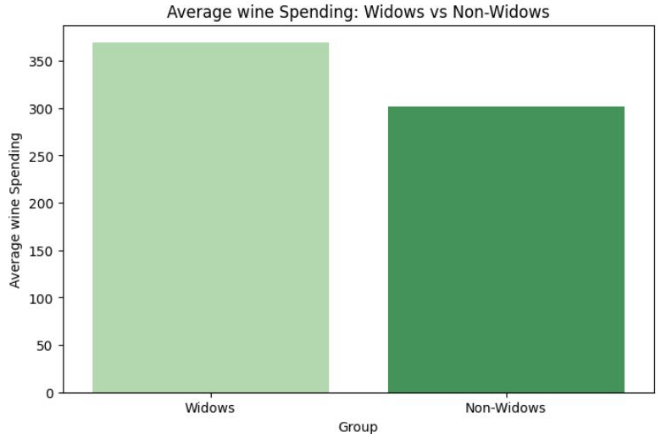
➔ Widows may represent a higher-value segment in terms of spending potential for wine.

➔ Marketing strategies could be tailored to target widows with premium or higher-value wine offers.

T-statistic: 1.740881806021484
P-value: 0.0836927457139911
Fail to reject the null hypothesis: No significant difference in wine spending

The statistically insignificant p-value to conclude : We cannot conclude from this data that widows spend more on wine





Hypothesis Testing - Time for some fun insights!

Widows are spending more?

What about gold?

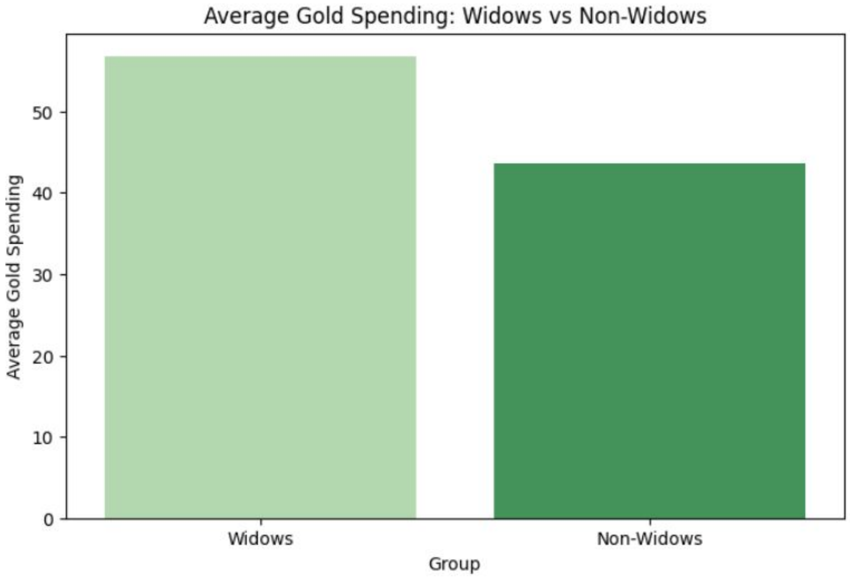
➔ Widows have a higher average monetary spending on gold but is it a coincidence?

Implications for Marketing:

➔ Widows may represent a higher-value segment in terms of spending potential for gold.
➔ Marketing strategies could be tailored to target widows with premium or higher-value wine offers.

T-statistic: 2.117253283106048
P-value: 0.03730048648313476
Reject the null hypothesis: Widows tend to spend more on gold

The statistically significant p-value to conclude : Widows do spend more on gold
Market re-entry? Well… we don't have a hypothesis for that, future scope maybe



Hypothesis Testing - Time for some fun insights!

What about the old and single people?

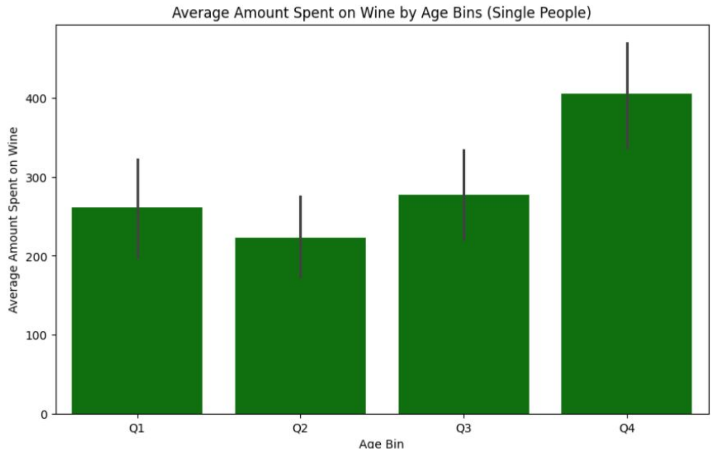
Single people tending to spend more on wine as they grow older or coincidence?

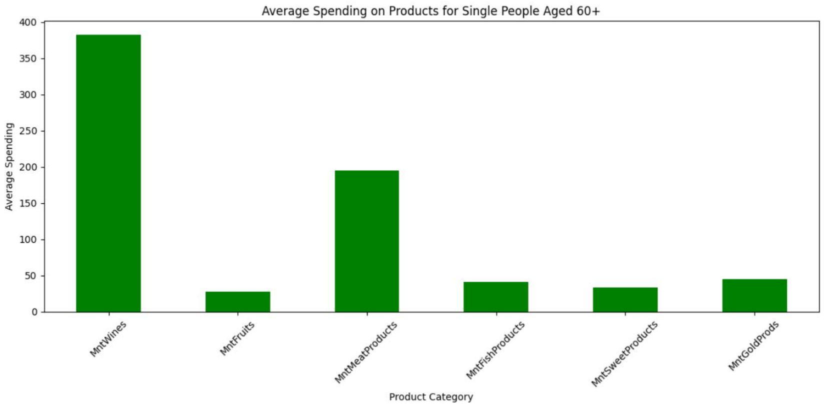
Implications for Marketing:

- ➔ Singles may represent a higher-value segment in terms of spending potential for wine
- ➔ Marketing strategies could be tailored to target singles with premium or higher-value wine offers for older customers

T-statistic: 2.500405948925115
P-value: 0.012476122072756477
Reject the null hypothesis: There is a statistically significant difference in wine spending between old single people and others.

The statistically significant p-value to conclude : Old single people actually tend to spend more on wine





Defining our high level strategy!

Approach 1

Vanilla model - Baseline model without any feature engineering

Approach 2

Define a RFM score to determine high-value customer score and then standardize. After standardizing, consider only +ve values as high value customers and use evaluation score to train the model by dropping features used for calculating score

Approach 3

Defining the same high value customers as above but training the model using all dropped features from above and this time dropping the evaluated score

Strategy to Predict Customer Campaign Response

Demographic features

+

RFM Score

+

Past campaign Results

Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer
1957	Graduatioı	Single	58138	0	0	4/9/12
1954	Graduatioı	Single	46344	1	1	8/3/14
1965	Graduatioı	Together	71613	0	0	21-08-2013
1984	Graduatioı	Together	26646	1	0	10/2/14
1981	PhD	Married	58293	1	0	19-01-2014

Requires a unique approach for effective segmentation of high-value customers

AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

RFM Analysis

After running univariate and bivariate analysis, we identify that Frequency and Monetary contribute significantly to what we define as a high value customer

Attributes taken

3

RECENCY ANALYSIS

- ➔ Recency appears to be pretty uniform across the range of values, with most customers having a recency between 0 and 100 days
- ➔ Customer base likely to have equal distribution of mix of recently active and somewhat inactive customers

Recency

FREQUENCY ANALYSIS

- ➔ Distribution is right-skewed indicating most of the customers are making fewer than 10 purchases
- ➔ Smaller customer segment would make multiple purchases, while most of them have lower engagement

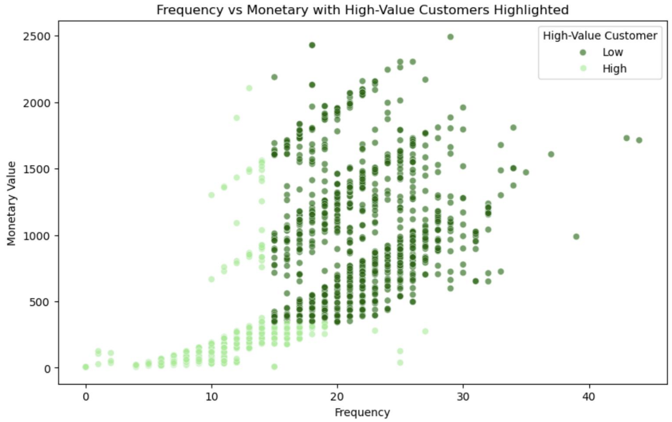
NumDealsPurchases
NumWebPurchases
NumCatalogPurchases
NumStorePurchases
NumWebVisitsMonth

MONETARY ANALYSIS

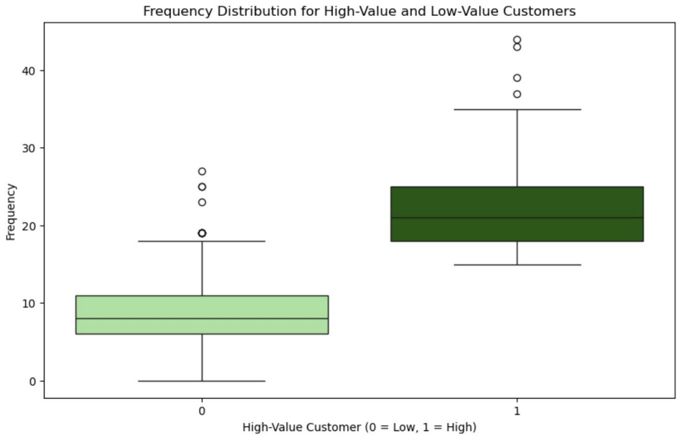
- ➔ This distribution heavily right skewed indicating that most of the customers are spending low amounts, while a very few have high spending levels
- ➔ Customer segment - low-value customers with small segment of high-value customers

MntWines
MntFruits
MntMeatProducts
MntFishProducts
MntSweetProducts
MntGoldProds

Method for Calculating Customer Scores



Positive correlation between Frequency and Monetary value indicates that customers who purchase more frequently tend to spend more, making these metrics key indicators for identifying high-value customers



High-value customers generally exhibit higher transaction frequencies than **low-value customers,** with a wider range and some outliers indicating particularly frequent buyers

High-value customers have a significantly higher median monetary value than **low-value customers** with a broader distribution and several outliers representing substantial spenders

RFM Score Calculation

$$\text{RFM score} = \alpha + \beta_1 R + \beta_2 F + \beta_3 M + \beta_4 (R \cdot F) + \beta_5 (F \cdot M) + \beta_6 (M \cdot R) + \beta_7 (R \cdot F \cdot M)$$

To determine the values of α and β coefficients, we use a regression model. Running this model requires inputs for both X & Y

- ➔ *X* - R, F, and their interaction terms R · F, M.F ,M.R, R · F · M, (using M as the base term)
- ➔ *Y* - whether a customer is high-value (1) or not (0)

Defining High-Value Customers

A high-value customer is identified if:

- ***Frequency > threshold of frequency***
- ***Monetary > threshold of monetary***

Here, we set the threshold for both Frequency and Monetary as the 75th quartile (0.75). High-value customers are represented by 1 in the results, while low-value customers are represented by 0.

RFM Score - Results from Logistic Regression Analysis

Accuracy Findings:

- ➔ With a **threshold of 0.75** for both Frequency and Monetary, the logistic regression model achieved an **accuracy of 93%** in identifying high-value customers.
- ➔ Adjusting the thresholds results in higher accuracy:
  - ◆ For Frequency = 80 and Monetary = 60, accuracy changes to 90%.
  - ◆ For Frequency = 60 and Monetary = 85, accuracy reaches 88%.

Resulting Coefficients:

The logistic regression model yielded the following values for α and β:

```
alpha = -0.024235447654728647
beta1 (Recency) = -0.14134785016010032, beta2 (Frequency) = -0.17865741479346944
beta4 (Recency * Frequency) = -0.0019375330209376528
beta5 (Frequency * Monetary) = 0.00012979334682973982
beta6 (Monetary * Recency) = -0.0002934186806590266
beta7 (Recency * Frequency * Monetary) = 1.9958562651709736e-05
```

The calculated α and β values were applied to the RFM formula to generate scores, which were then used to improve customer segmentation accuracy in the final model

Results from performing different models

Approach 1

Approach 2

Approach 3

Model Type	Class	Precision	Recall	F-1 Score
Logistic Regression	0	0.86	0.97	0.91
	1	0.45	0.13	**0.20**
Random Forest	0	0.89	0.98	0.93
	1	0.72	0.33	**0.46**
XGBoost	0	0.89	0.96	0.93
	1	0.63	0.38	**0.47**
Neural Network (yes, we know it's an overkill)	0	0.90	0.94	0.92
	1	0.56	0.42	**0.48**

Results from performing different models

Approach 1

Approach 2

Approach 3

Model Type	Class	Precision	Recall	F-1 Score
Logistic Regression	0	0.84	0.94	0.89
	1	0.62	0.34	**0.44**
Random Forest	0	0.84	0.91	0.88
	1	0.55	0.38	**0.45**
XGBoost	0	0.87	0.91	0.89
	1	0.62	0.51	**0.56**
Neural Network (yes, we know it's an overkill)	0	0.87	0.92	0.90
	1	0.65	0.51	**0.57**

Results from performing different models

Approach 1

Approach 2

Approach 3

Model Type	Class	Precision	Recall	F-1 Score
Logistic Regression	0	0.80	0.94	0.86
	1	0.35	0.13	**0.19**
Random Forest	0	0.84`	0.92	0.88
	1	0.57	0.36	**0.44**
XGBoost	0	0.88	0.90	0.89
	1	0.58	0.53	**0.56**
Neural Network (yes, we know it's an overkill)	0	0.89	0.87	0.88
	1	0.56	0.60	**0.58**

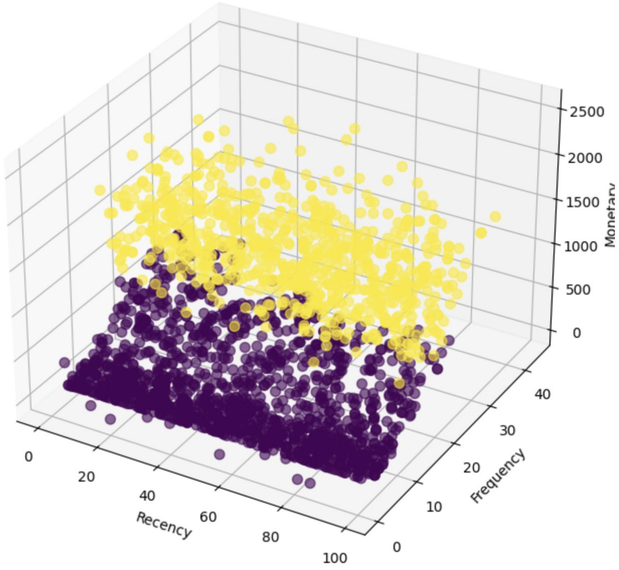
Unsupervised Model: K-Means Clustering

We applied K-Means, to assess its effectiveness in predicting high-value customers using RFM and then run the classification models to check how well it's performing

Clustering Results:

Number of clusters = 2 *(Since we are doing high value and not high value customers)*

	Recency		Frequency		Monetary		RFM_Score
	mean	median	mean	median	mean	median	mean
Cluster							
0	48.057840	48.0	10.921951	9.0	209.294774	101.0	268.274564
1	50.983851	54.0	21.885714	22.0	1312.608696	1229.0	1385.478261



We considered cluster 1 as the high value customers since it has the highest RFM mean score and for further analysis by running classification models on it

Unsupervised Model - K means Clustering Results for Classification

Model Type	Class	Precision	Recall	F-1 Score
Logistic Regression	0	0.83	0.94	0.88
	1	0.61	0.31	**0.41**
Random Forest	0	0.86	0.97	0.91
	1	0.80	0.44	**0.57**
XGBoost	0	0.88	0.92	0.90
	1	0.67	0.56	**0.61**
Neural Network (yes, we know it's an overkill)	0	0.89	0.88	0.88
	1	0.65	0.68	**0.67**

Final Prediction Results & Key Insights

Do high-value customers actually respond higher to campaign than others? What does our analysis say?

Approach 1- Vanilla model:

Using all the data features from our raw dataset and fitting in the models resulted in **“bad decision making”**

Approach 2 - RFM Score:

Predicting on features with RFM metric and demographic features post segmentation using engineered evaluation score gave us better results than approach 1

Approach 3 - Predictions on features without Score:

Predicting on features without RFM metric and demographic features also gave us similar accuracy as approach 2

Clustering - how unsupervised made all the difference!

K-means Clustering resulted in the most accurate predictions compared to the supervised approaches

Thank You!

We know we are not just between you and your weekend this time – rather we're standing in the way of your turkey dreams and Thanksgiving feasts!