Predict Customer Personality to boost marketing campaign by using Machine Learning

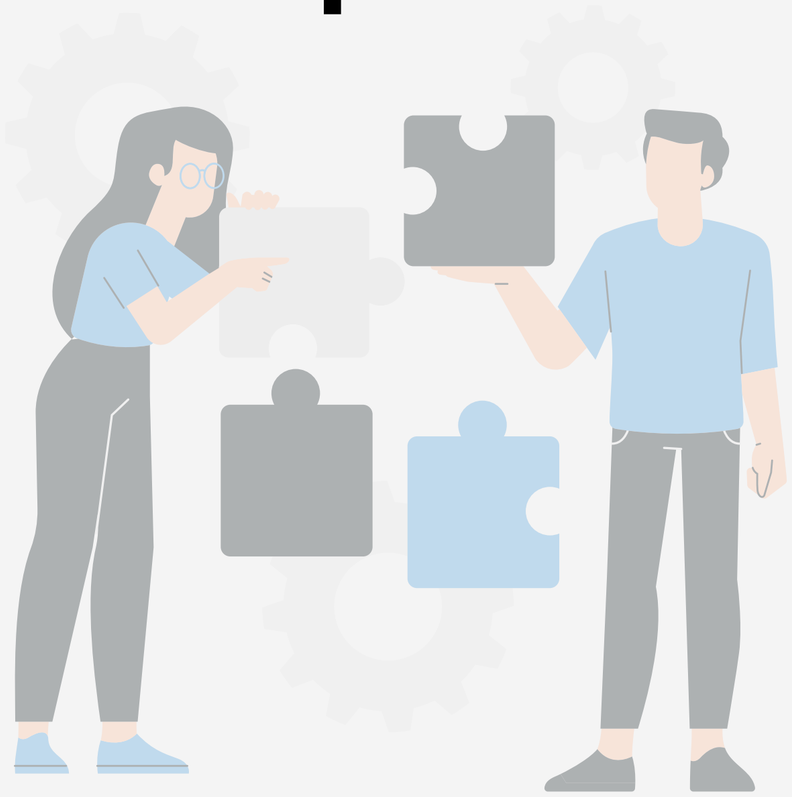
Created by:
Lutfia Husna Khoirunnisa
Your Email:
lutfiahusnakhoirunnisa@gmail.com
Your linkedIn Profile
.linkedin.com/in/lutfiahusnakhoirunnisa

“Lutfi is a junior data analyst experienced on data analysis, business analysis, and data science with a background in mathematics. Experienced in handling and interpreting diverse data sets, extracting valuable insights, and making datadriven recommendations ”

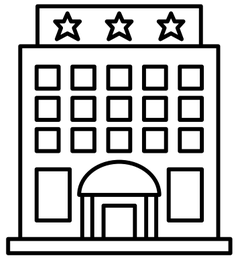
Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Rakamin Academy

Step by Step



1. Project Introduction
2. Problem Statement
3. Data Overview and Preprocessing
4. Data Modeling
5. Conclusion



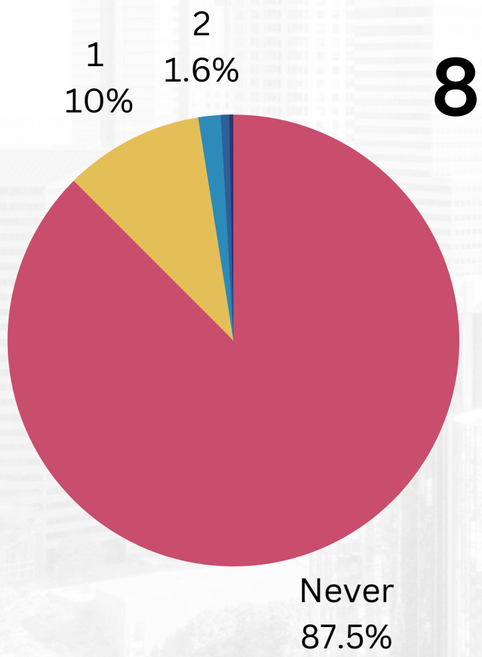
Project Introduction

A company can thrive when it knows the behavior and personality of its customers, allowing it to offer better services and benefits to potential loyal customers. By analyzing past marketing campaigns to improve performance and target the right customers on the company's platform, our main goal is to create a predictive clustering model that helps the company make smarter decisions

All data model and visualization in this project are built with Python

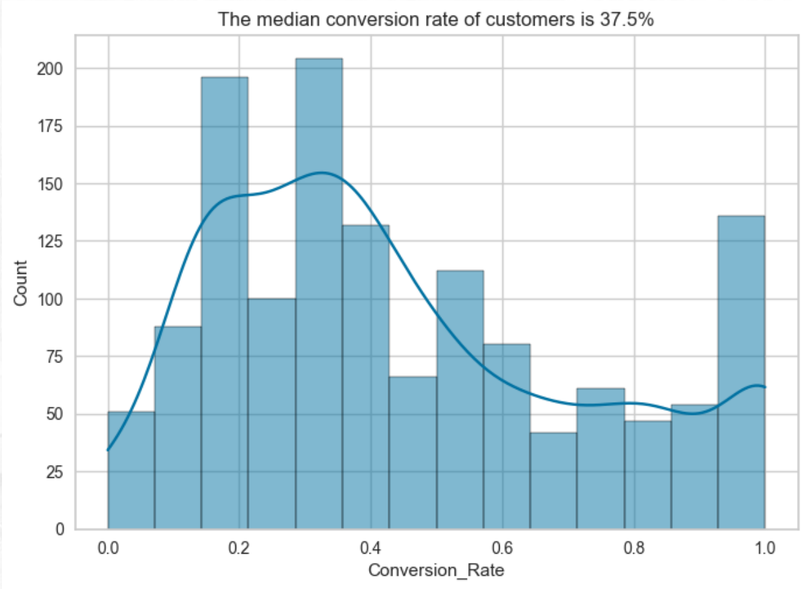
access the dataset and notebook here

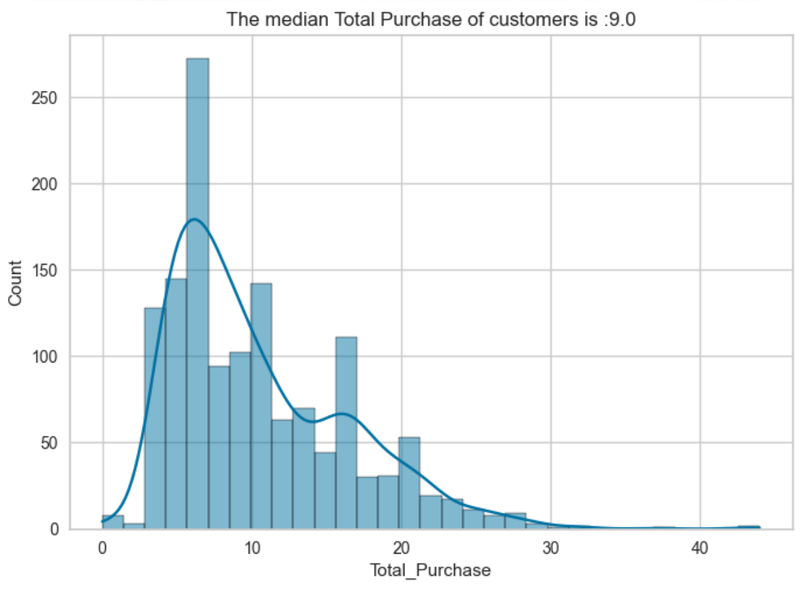
Problem Statement



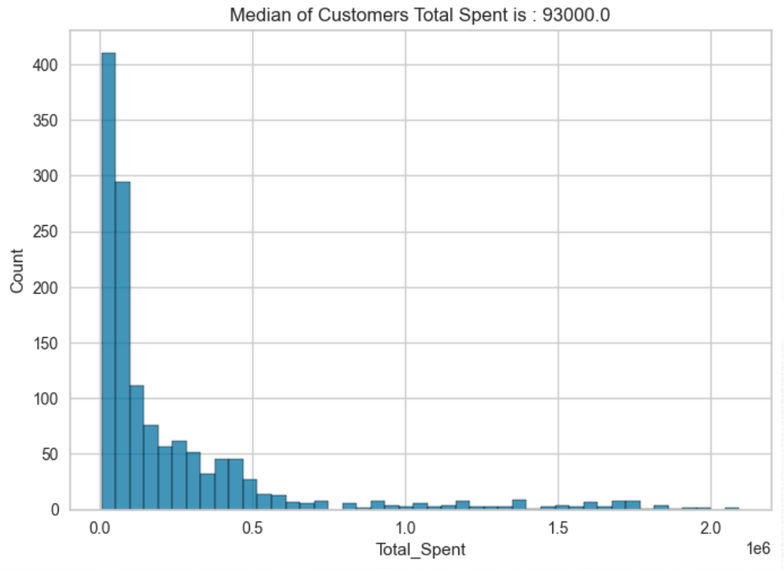
87.5% of customers have never accepted any campaigns given by the company.

To increase sales, the company can maximize campaigns by providing right campaigns to each customer. This can help improve the low conversion rate and attract customers to spend more



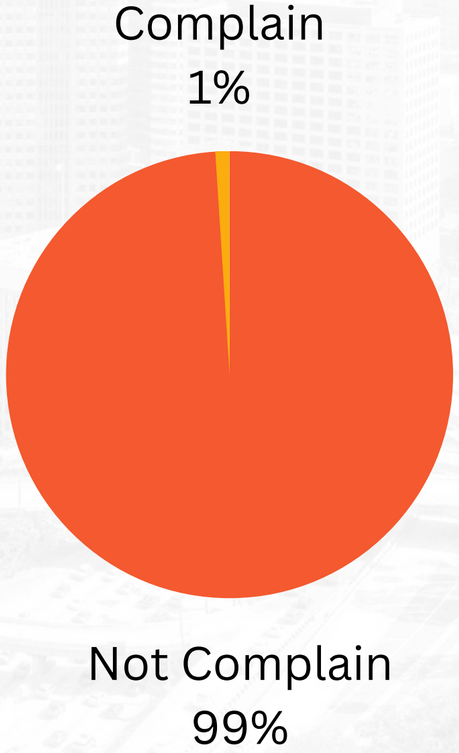


Problem Statement



From the total customer spend distribution, we know that the majority of customers have low total spend. The customer conversion rate distribution also indicate that the majority of customers have low values of conversion rate.

As we know, 99% of customers have never complained in the past 2 years. The low total spend and conversion rate of customers may not be caused by poor service or bad customer experience. So, this fact indicate that the campaigns given may not have effectively reached the customers, causing them to feel less interested in making a purchase.



Problem Statement

"The company needs to be focused on attracting customers to make transactions. If the company wants to focus on strategy of providing campaigns to customers, the marketing team needs to evaluated and improve their old strategies, especially considering that the majority of customers have never received any campaigns at all"

Problem Statement

Rakamin Academy

Role

Data Analyst

Business Metrics

TOTAL ACCEPTED CAMPAIGN

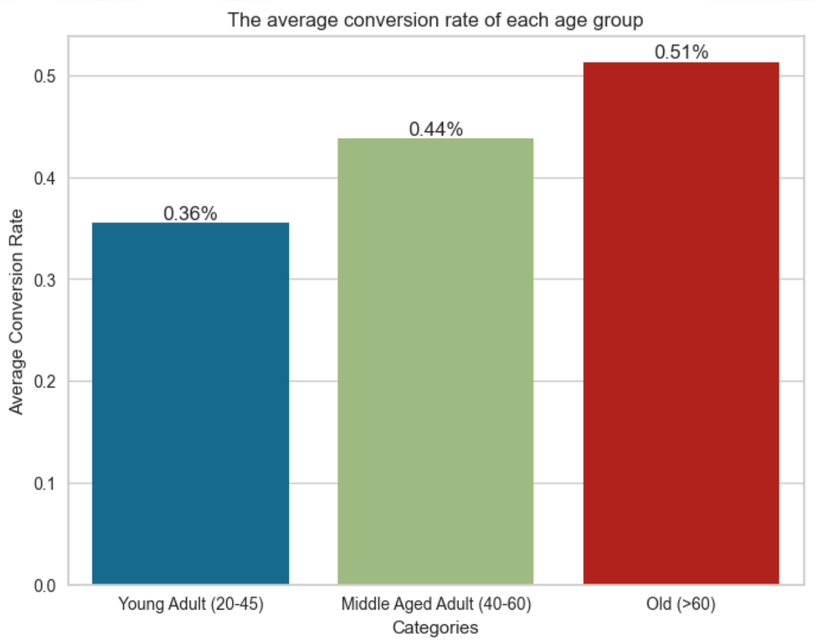
Goal

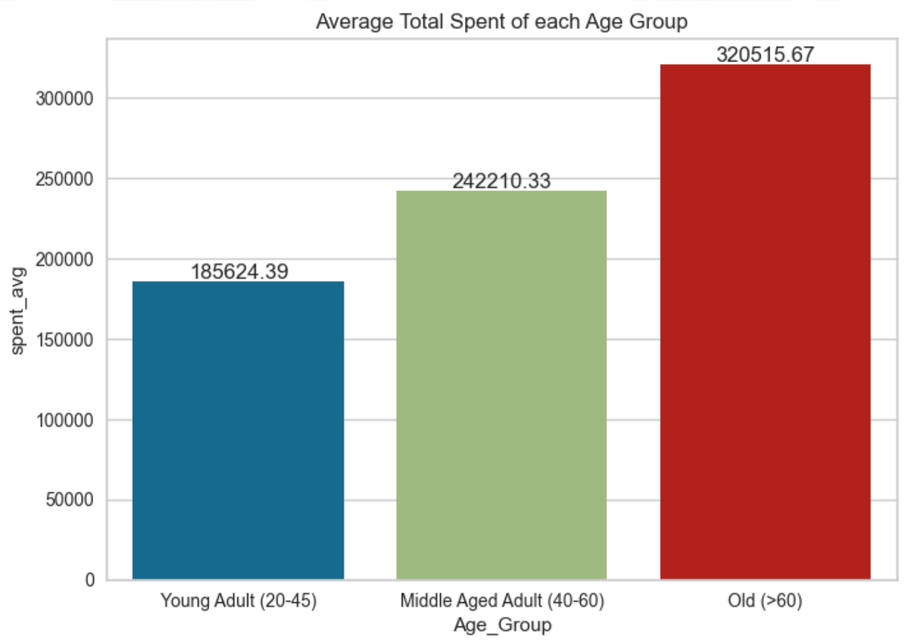
Increase the total number of campaigns accepted by customers.

Solution

Creating a model to perform customer segmentation using unsupervised learning.

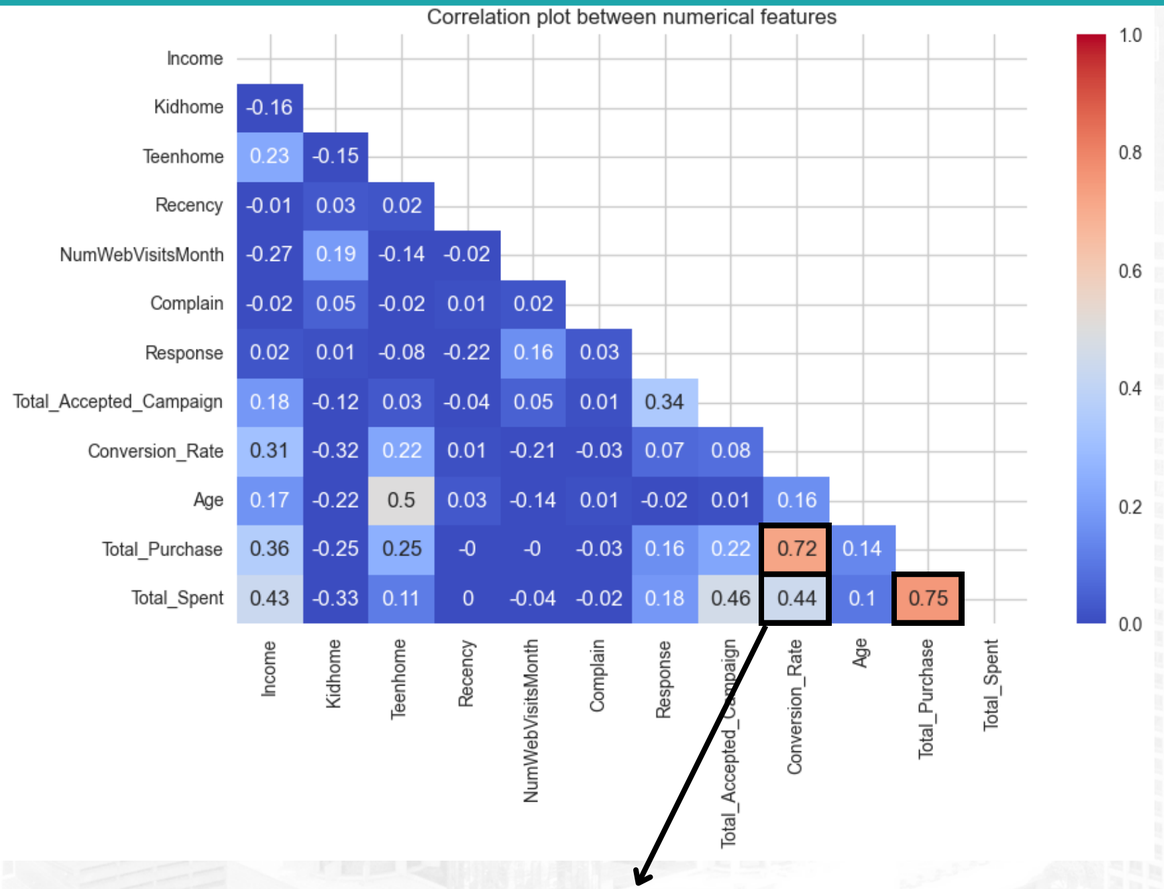
Problem Statement
(customers overview)





The 'old' age group has the highest average conversion rate and total spent among the other age groups. If the company wants to focus on increasing sales, they can prioritize offering a promo or discount that provides the needs and preferences of this age group.

Problem Statement
(customers overview)

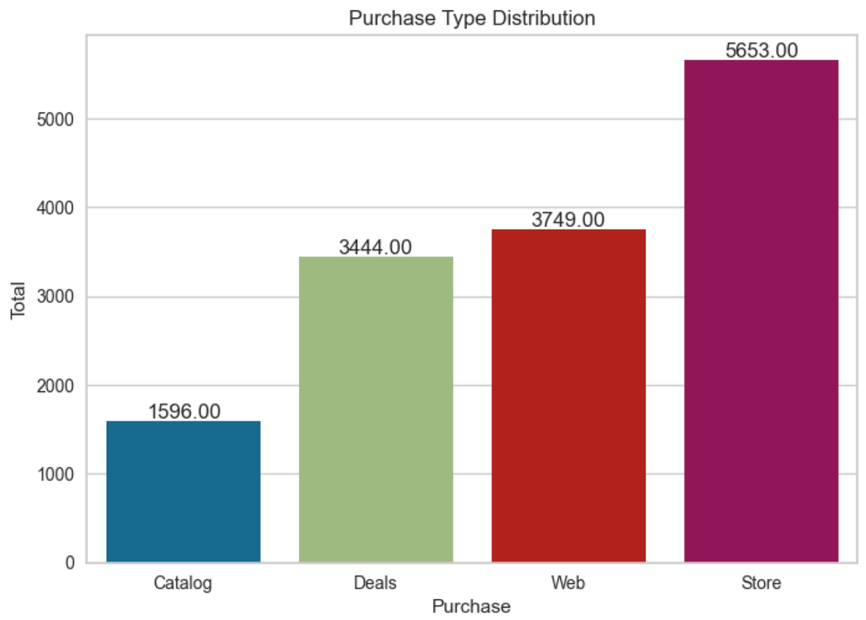


From the heatmap, there are some fact that we need to poit out :

- Total Purchase have high correlation with Total Spent, and conversion rate.
- Income does not have a strong correlation with total spent or conversion rate, so income does not influence these features.

- Total Spent has a moderate correlation with Total Spent, indicating that many customers may prefer to make transactions through platforms other than the website.

From the graph on the side, we can see that customers prefer making transactions in-store rather than on the web.



Data Overview & Preprocessing

This project will use a dataset that contains **customers data historical marketing campaign**, including the number of customers, age, income, total spend, and more. The dataset consists of 2240 records and 30 features.

The dataset is not yet clean as it contains missing values, incorrect values, and unused records that require handling

DATA CLEANSING :

- Handling missing value in the 'Income' column by imputate it wuth median value
- Droping unused column such as Z_Revenue, Z_CostContact column, and 'Unnamed: 0' columns
- Droping incorrect data, such as customerd data with number of web activity lesss than web purchase,

DATA PREPROCESSING

Feature selection, outlier handling, and feature normalization also will be conducted

FEATURE ENGINEERING

Create several new columns:

- Total_Purchase: represents the number of purchases made by customers.
- Total_Spent: indicates the total amount spent by customers.
- Total_Accepted_Campaign: denotes the number of campaigns accepted by each customer.
- Conversion_Rate: measures the percentage of users or visitors who take a desired action.

access the dataset and notebook here

Data Overview & Preprocessing

Rakamin Academy

FEATURE SELECTION :

In this clustering, we will select features that are considered to represent customers well. **The RFM framework** will be used to define which features are used to represent customers. The RFM framework is a customer segmentation framework that combines multiple factors to categorize customers into distinct groups. Each letter in RFM represents a specific aspect of customer behavior:

R

Recency measures the number of days that have elapsed since the customer's last purchase, and we will use the 'Recency' column for this.

F

Frequency indicates the number of transactions that the customer has made, and we will use the 'Total_Purchase' column for this.

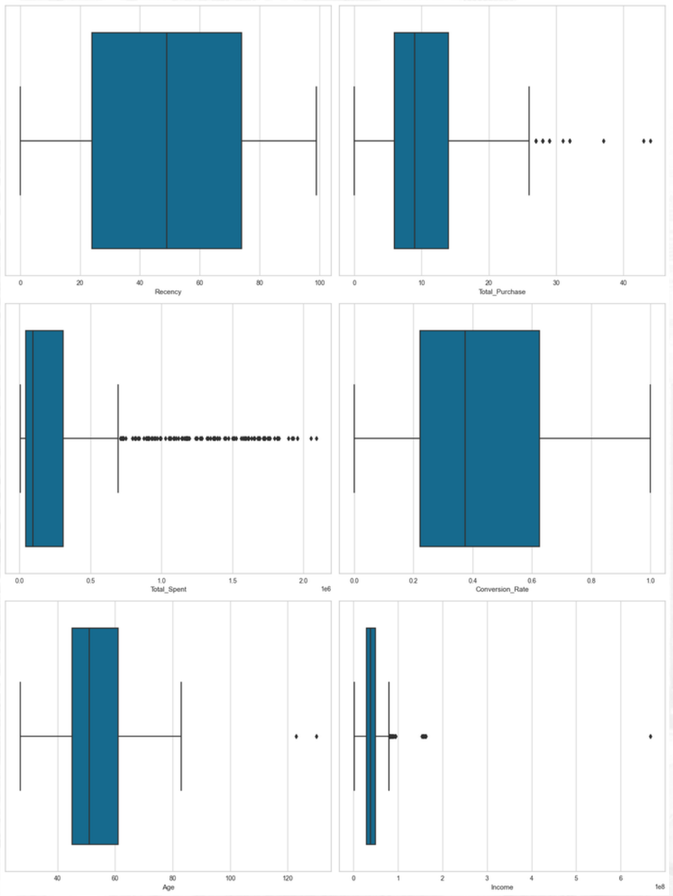
M

Monetary value represents the amount of money that the customer has spent, and we will use the 'Total_Spent' column for this.

+

In addition to these features, we will also consider additional features:
Demographic segmentation involves categorizing customers based on demographic information, and we will use the 'Age' column for this.
Behavioral segmentation involves categorizing customers based on their behavior, and we will use the 'conversion rate', and 'Income' columns for this.

Data Overview & Preprocessing



OUTLIER HANDLING

Since K-means clustering will be used, which is based on distance calculation, the data considered as outliers will be separated from the overall data to avoid compromising performance and ensure better cluster formation. Afterwards, the outliers will be assigned to their respective clusters using the K-nearest neighbors (KNN) classification method, which also relies on distance calculation.

From the conducted data analysis, the following determinations are made:

- Customers with more than 30 purchases are considered outliers.
- Customers who older than 100 years old are considered outliers.
- Customers with more than 14000000 worth of income are considered outliers.
- Customers with more than 2,000,000 worth of spent are considered outliers.

And we got 1353 rows of data that will be used, and 16 rows of outliers data

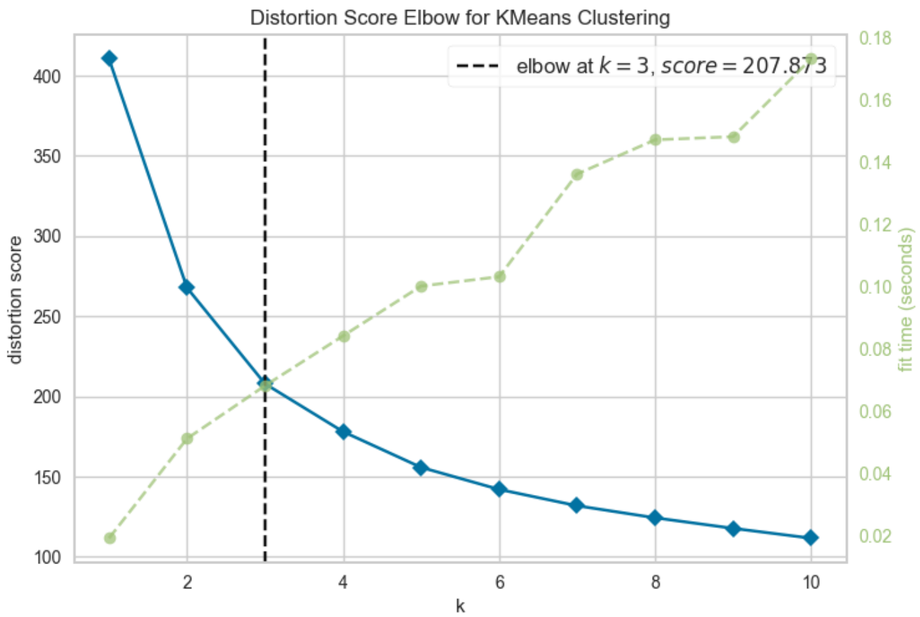
FEATURE NORMALIZATION

Normalization is applied to all features except for the conversion rate feature. As the k-means method relies on distance calculations, it is necessary to scale the data to prevent biases. The conversion rate feature is excluded from normalization since it is already normally distributed and represented as percentages, sharing the same scale.

Data Modeling

Rakamin Academy

To determine the optimal number of clusters or groups for data segmentation, an elbow test and silhouette test will be conducted. These tests will help identify the ideal number of customer segments.



According to the elbow test, the optimal number of clusters for data segmentation in this dataset is 3.

Then, the clustering model using Kmeans and 3 clusters applied to the non-outlier data

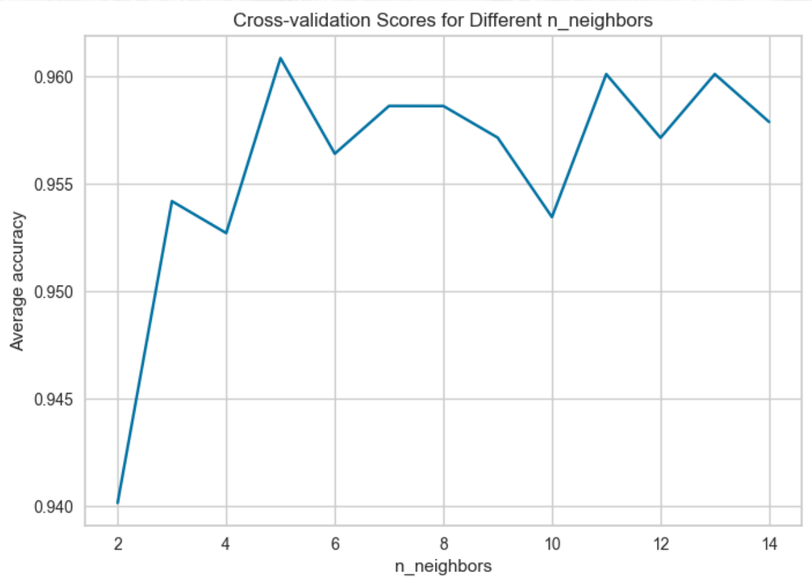
```
# Clustering model
kmeans = KMeans(n_clusters = 3, init = 'k-means++', n_init = 10, random_state = 0)
kmeans.fit(model_scaled)
model['CLUSTER'] = kmeans.labels_
model_scaled['CLUSTER'] = kmeans.labels_
```

The silhouette score obtained for this clustering is 0.5985

Data Modeling

The next step is determining the cluster for the outlier data using the KNN classification method. This algorithm is selected due to its reliance on distance calculations, making it well-suited for cluster assignment.

In order to determine the optimal number of neighbors for classification with the KNN algorithm, a search will be conducted using cross-validation on the training data derived from the clustering dataset.



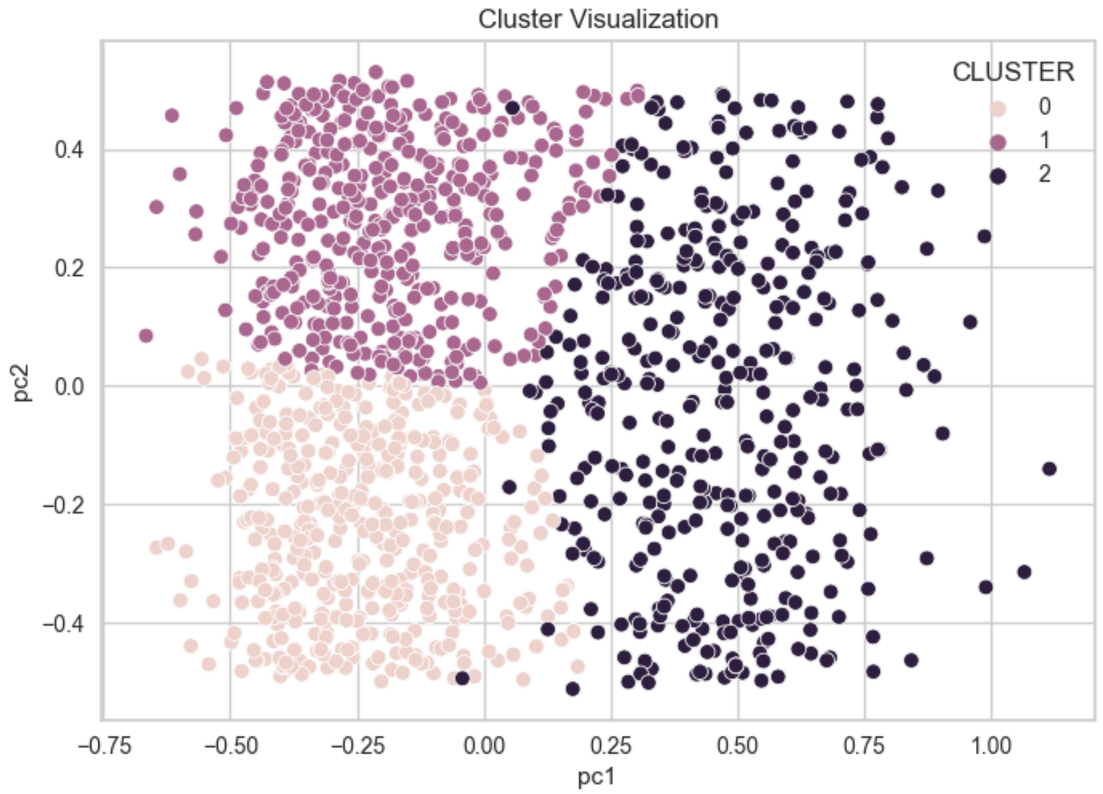
It is found that n_neighbors = 5 is sufficient for classification in this data.

```
# create KNN model with k=3
knn = KNeighborsClassifier(n_neighbors=5)

# fit the model using the training data
knn.fit(model_scaled.drop('CLUSTER', axis = 1), model_scaled['CLUSTER'])

# predict on the test data
outlier['CLUSTER'] = knn.predict(outlier_scaled)
```

Data Modeling

Rakamin Academy

From the obtained visualization, it can be observed that the data is well separated based on their clusters.



Conclusion

Rakamin Academy

From the obtained clusters, there are 3 different customer segments

CLUSTER	Recency (mean)	Total_Purchase (med)	Total_Spent (med)	Conversion Rate (med)	Age (med)	Income (med)
0	24.80	7.0	55000.0	0.29	49.0	32570000.0
1	76.01	7.0	62000.0	0.29	51.0	34469000.0
2	47.19	17.0	428500.0	0.80	54.5	51991000.0

CLUSTER 0 - The Newer Customers

These customers are new with low total spent, conversion rate, and purchase frequency. They tend to be younger customers with low income.

CLUSTER 1 - The Inactive customers

Based on their recency, these customers are long-time customers who have not made any transactions. They have relatively low conversion rates and number of purchases, with medium spending. These customers are typically adults with moderate income.

CLUSTER 2 - The Loyal Customers

Based on their recency, these customers are still actively making transactions. Customers in this group have high conversion rates, number of purchases, and total spending. They tend to be relatively older customers with high income.

Conclusion

BUSINESS RECOMMENDATION

CLUSTER 0 - The Newer Customers

For the newer customers, the company needs to implement strategies to **attract and retain** these customers for repeat transactions. This can be achieved by offering promotions such as discount for their next transaction, or loyalty programs with exclusive benefits

CLUSTER 1 - The Inactive customers

For the inactive customers, the company should focus on strategies to **re-engage** them and encourage them to make transactions again. This can include offering exclusive special promotions or discounts for returning customers, as well as personalized offers tailored to their previous preferences. It is also important to conduct customer experience analysis to understand the reasons behind their disengagement and identify what to improve.

CLUSTER 2 - The Loyal Customers

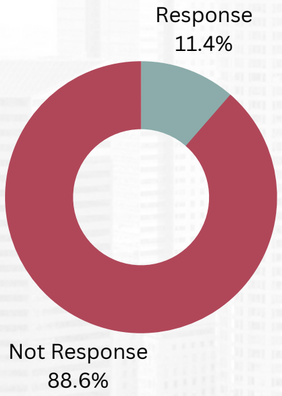
For the Loyal Customers, the company should focus on strategies aimed at **maintaining** customer loyalty, such as offering rewards, customer discounts, or more enticing promotions.

Conclusion

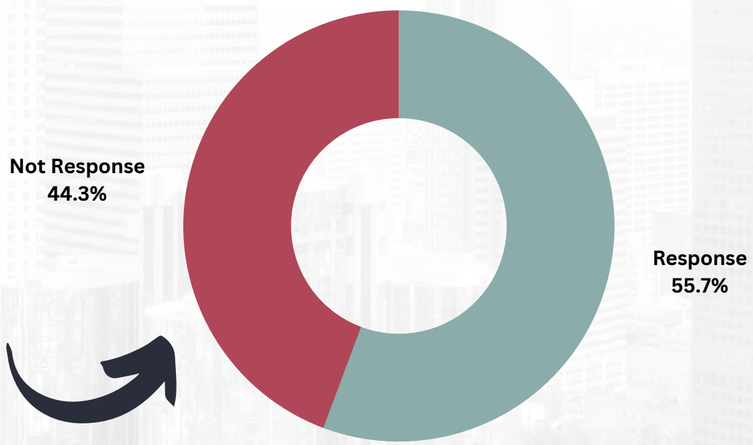
Potential Impact

In order to assess the impact, we utilize customer response data from the most recent campaign.

Before the model



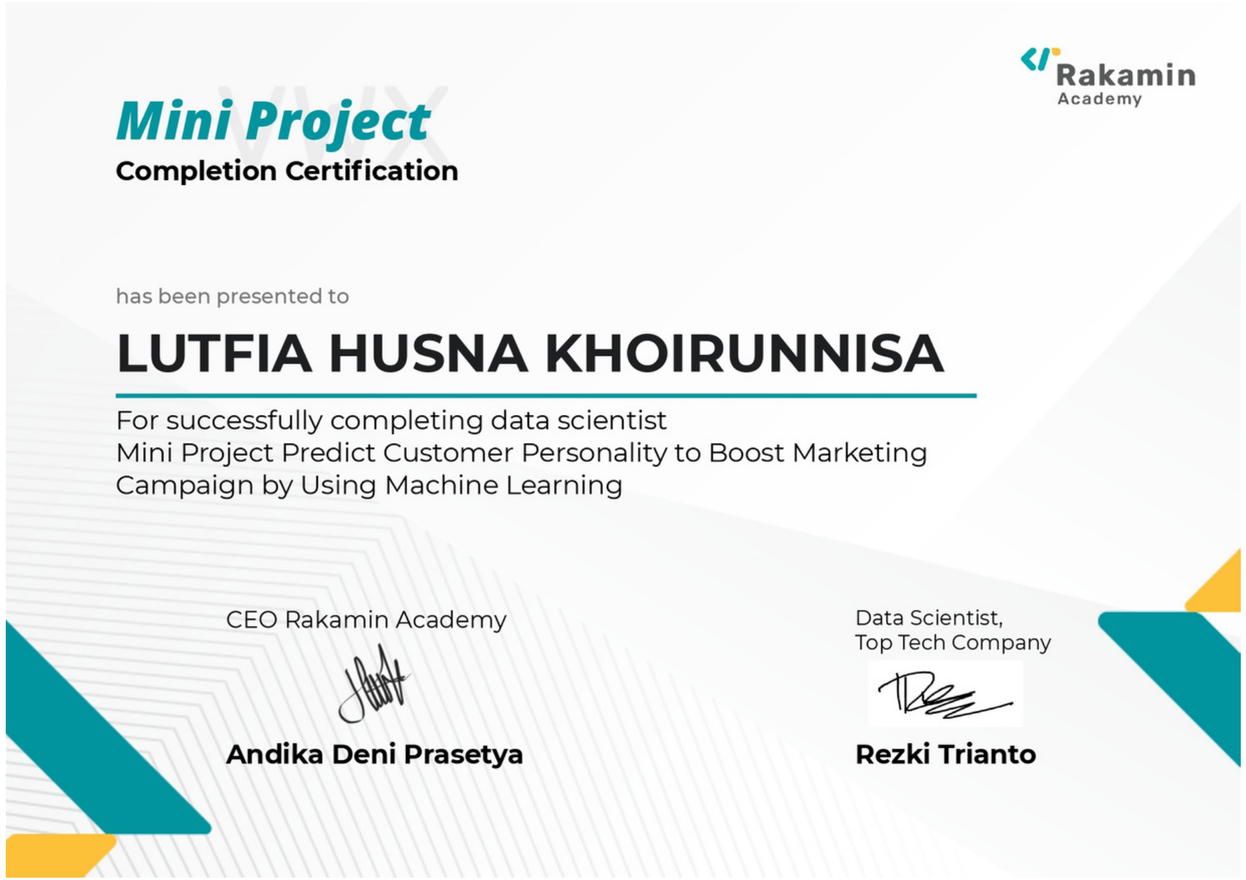
After the model



Assuming that customizing the campaign for each customer segment can convert half of the non-responsive customers into responsive ones, we can achieve a four-fold increase in customer response compared to the pre-implementation of campaign segmentation on customers.

To calculate the potential impact, disregarding other factors and new customers, we assume that only customers who respond to the campaign make transactions. With the model and its implementation, we can potentially achieve a Gross Merchandise Value (GMV) of around **95,7M - 321,M in the next year.** Whereas, without the model, the GMV obtained is only around 69,3M.

So, by implementing personalized campaigns for each customer segmentation, the GMV can be increased by approximately **1 to 4 times.**



Thank you

Check out my profile and other portfolios:

Linkedin

GitHub