Predict Customer Personality to boost marketing campaign by using Machine Learning

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

Created by:
Raffel Ravionaldo
raffelrazor@gmail.com
https://www.linkedin.com/in/raffel-ravionaldo/

A fresh graduate interested in data, learns about data through 2 data science bootcamp, first organized by Rakamin and the second by binar academy. To deepen my knowledge at data, I took part in a virtual intership by rakamin cooperating with ID/X Partner and Home Credit Indonesia, and a virtual project organized by forage.com cooperating with British Airways.

Overview

“ A company can develop rapidly when it knows the behavior of it’s customer personality, so that it can provide better services and benefits to customers who have the potential to become loyal customers. By processing historical marketing campaign data to improve performance and target the right customers, so they can transcat on the company’s platform, from this data insight our focus is to create a cluster prediction model to make it easir for companies to make decisions.

Dataset

```
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   ID                   2240 non-null   int64  
 1   Year_Birth           2240 non-null   int64  
 2   Education            2240 non-null   object 
 3   Marital_Status       2240 non-null   object 
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64  
 6   Teenhome             2240 non-null   int64  
 7   Dt_Customer          2240 non-null   object 
 8   Recency              2240 non-null   int64  
 9   MntCoke              2240 non-null   int64  
 10  MntFruits            2240 non-null   int64  
 11  MntMeatProducts      2240 non-null   int64  
 12  MntFishProducts      2240 non-null   int64  
 13  MntSweetProducts     2240 non-null   int64  
 14  MntGoldProds         2240 non-null   int64  
 15  NumDealsPurchases    2240 non-null   int64  
 16  NumWebPurchases      2240 non-null   int64  
 17  NumCatalogPurchases  2240 non-null   int64  
 18  NumStorePurchases    2240 non-null   int64  
 19  NumWebVisitsMonth    2240 non-null   int64  
 20  AcceptedCmp3         2240 non-null   int64  
 21  AcceptedCmp4         2240 non-null   int64  
 22  AcceptedCmp5         2240 non-null   int64  
 23  AcceptedCmp1         2240 non-null   int64  
 24  AcceptedCmp2         2240 non-null   int64  
 25  Complain             2240 non-null   int64  
 26  Z_CostContact        2240 non-null   int64  
 27  Z_Revenue            2240 non-null   int64  
 28  Response             2240 non-null   int64  
dtypes: float64(1), int64(25), object(3)
memory usage: 525.0+ KB
```

Description

Dataset contains customer data who made transcantions and interactions on our platform, such as total amount buying products at our shop, year of birth, is accepted our campaign or no and the number of purchases made from our store (either coming directly to the store, through the website etc)

Click here to see the code

Feature Engineering

Make a new column, there are :

No	Name of new column	How to do it?
1	Age	Substract this year (using datetime) with year birth of customers
2	Total Children	Sum of Kid Home and Teen home
3	Group of Age	From column age, if it’s < 35 we call it Young adult, 35 < x < 65 is adult and > 65 called Senior adult
4	Total of accepted campaign	Sum of accepted campaign at column that have ‘Acceptedcmp’ in columns name
5	Total of Purchases	Sum from column that have ‘Num’ at the column name (except NumWebVisitsMonth)
6.	Total amount	Sum from column that have ‘Mnt’ at the column name
7	Conversion rate	Total purchases divided by number of web visit per month
8.	Total Years joined	Count total day joined, after that divided it by 365

Click here to see the code

Exploratory Data Analysis

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

Total accepted campaign analysis

Rakamin Academy

	index	corr_matrix	dfbase	correlation
0	9	Total_accepted_campaign	Total_amount	0.459554
1	5	Total_accepted_campaign	Response	0.426035
2	0	Total_accepted_campaign	Income	0.307122
3	8	Total_accepted_campaign	Total_Purchases	0.257273
4	7	Total_accepted_campaign	total_children	0.244282

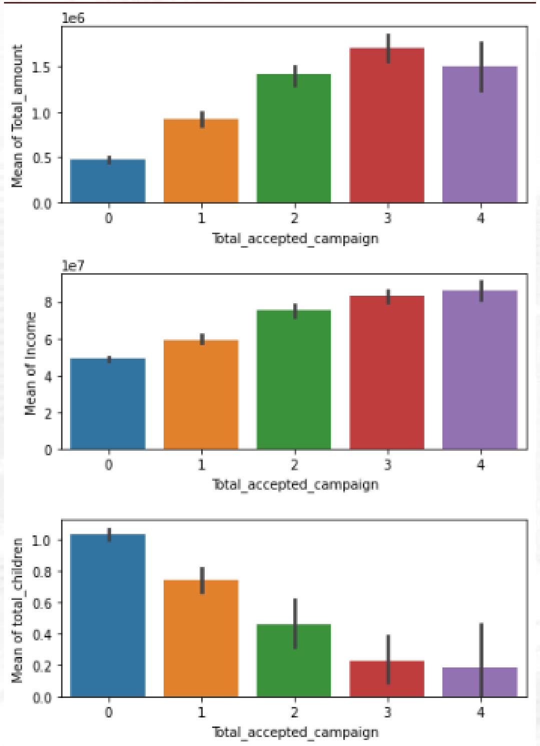
```
Chi-squared test for Education:
Chi-squared test statistic:  13.652661876664261
p-value:  0.6245722039336392


Chi-squared test for Marital_Status:
Chi-squared test statistic:  14.449720719561753
p-value:  0.8069820454432217


Chi-squared test for grup_age:
Chi-squared test statistic:  16.430737547135532
p-value:  0.036613853379265
```

Correlation between numeric column and total accepted campaign, we will analyst it only top 5 column, there are : 'Total_amount', 'Response', 'Income', 'Total_Purchases','total_children'

For categories column, we analyst just one column, there are grup_age because it have a lowest p-value at chi-square test between another categories column

Click here to see the code

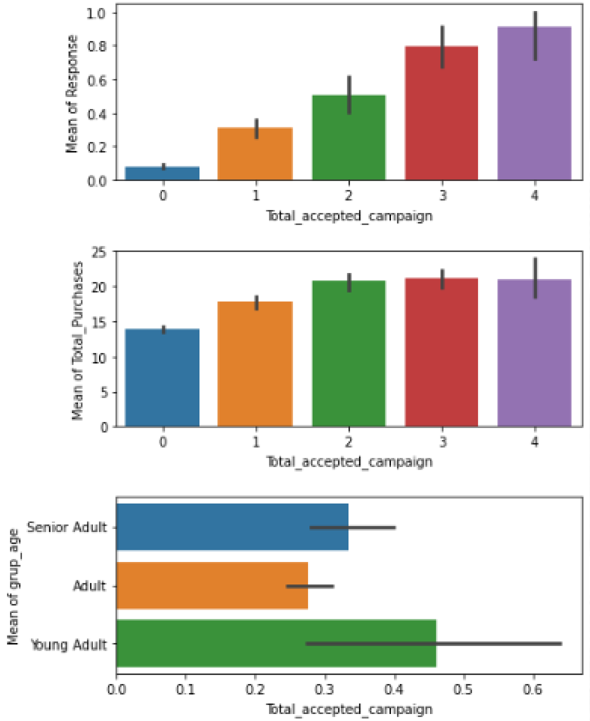
Total accepted campaign analysis



From picture beside, we can see that the higher total amount and income, then the higher number of total accepted campaign received and if the consumer have more children, then total_accepted_campaign will be lower.

[Click here to see the code]

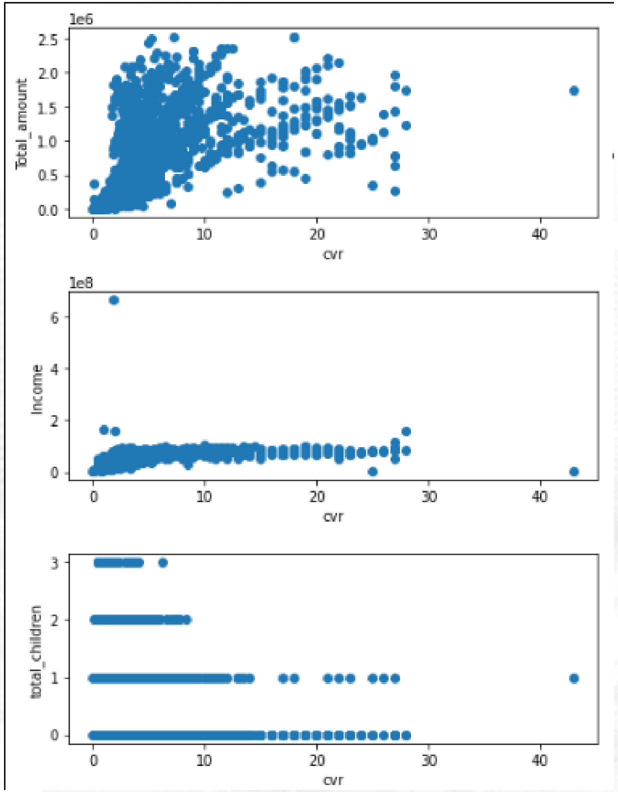
Total accepted campaign analysis



From picture beside, if customer have response our current campaign, the number of total_accepted_campaign will be higher, in total purchases, the higher puchases that the customer do, the higher number of total_accepted_campaign. And we can see Young adult have a tendency to accept the campaign that we provide.

Click here to see the code

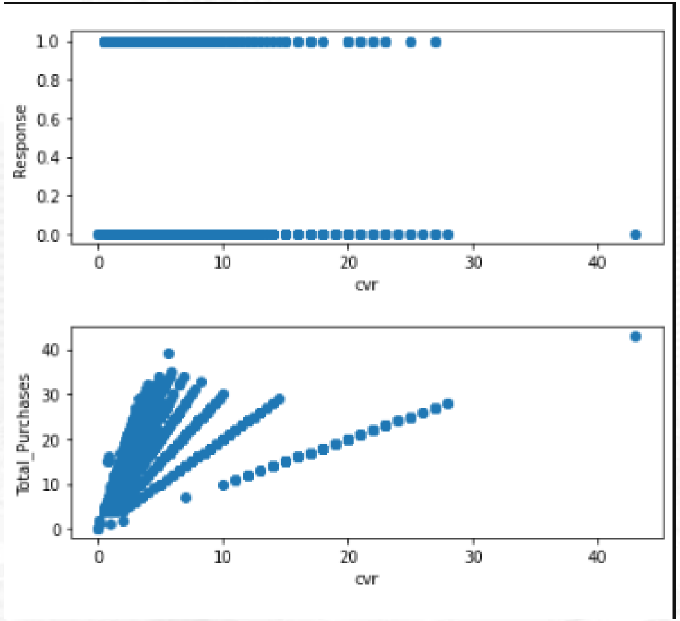
Conversion Rate analysis



From picture beside, we can see at total amount and income have a low positive correlation with cvr, at total children, if the customer have fewer children, the value of cvr will be higher.

Click here to see the code

Conversion Rate analysis



From picture beside, the distribution of data in the response column looks quite even, and at total purchase have a low positive correlation with cvr.

Click here to see the code

Data Preprocessing

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

Data Cleaning & Preprocessing

Rakamin Academy

index		Total Null Data
4	Income	24

At our data, only income column that have null value, so let's fill it with median (because income is numeric column)

```
df.duplicated().sum()

0
```

Our data don't have duplicated data, so we don't need to remove any rows.

```
not_marry = ['Lajang', 'Bertunangan']
marry = ['Menikah', 'Cerai', 'Janda', 'Duda']

marital_Status = []
for i in df['Marital_Status']:
    if i in not_marry:
        status = 'Never been married'
    else:
        status = 'Ever been married'
    marital_Status.append(status)

df['Marital_Status'] = marital_Status
```

At marital status, we change the value because in that column it has a value that has the same meaning

Click here to see the code

Data Cleaning & Preprocessing

Rakamin Academy

```
num_cmp = [col for col in df.columns if 'Num' in col]
mnt_cmp = [col for col in df.columns if 'Mnt' in col]

col_to_drop = num_cmp + mnt_cmp + acc_cmp

df.drop(columns = col_to_drop, inplace=True)
df.drop(columns = ['Year_Birth', 'ID', 'Kidhome', 'Teenhome'], inplace=True)
```

Drop unused column (mostly columns that have been used in feature engineering)

Click here to see the code

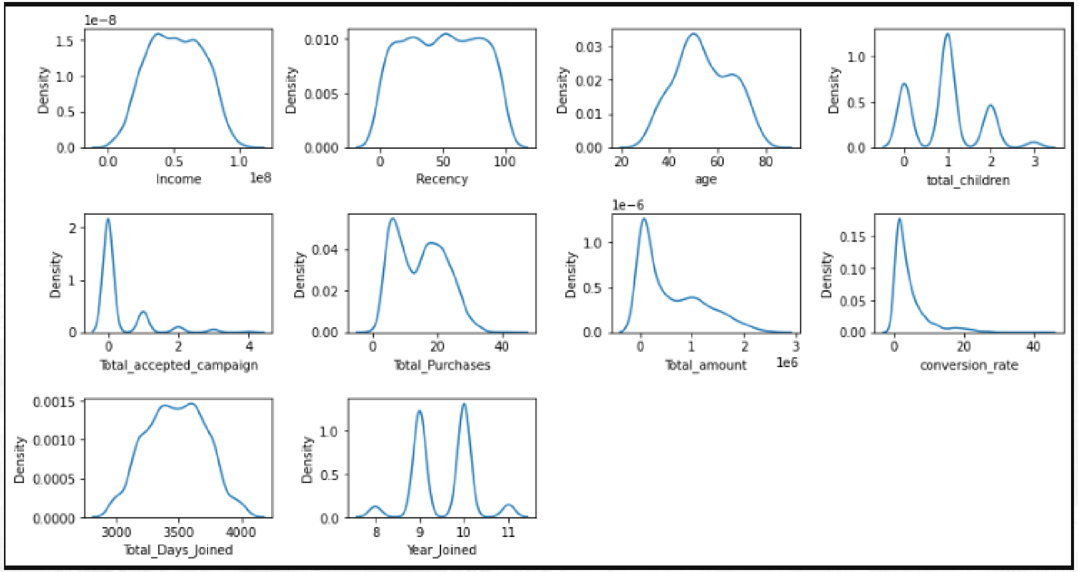
Feature Engineering

```
# Label Encoding
mapping_education = {
    'SMA' : 0,
    'D3'  : 1,
    'S1'  : 2,
    'S2'  : 3,
    'S3'  : 4
}

mapping_marital = {
    'Never been married' : 0,
    'Ever been married' : 1
}

mapping_grup_age = {
    'Young Adult' : 0,
    'Adult'  : 1,
    'Senior Adult' : 2
}

df['Education'] = df['Education'].map(mapping_education)
df['Marital_Status'] = df['Marital_Status'].map(mapping_marital)
df['grup_age'] = df['grup_age'].map(mapping_grup_age)
```

For category column, we do one-hot encoding because all the categories column have ordinal data.

Click here to see the code

Feature Transformation

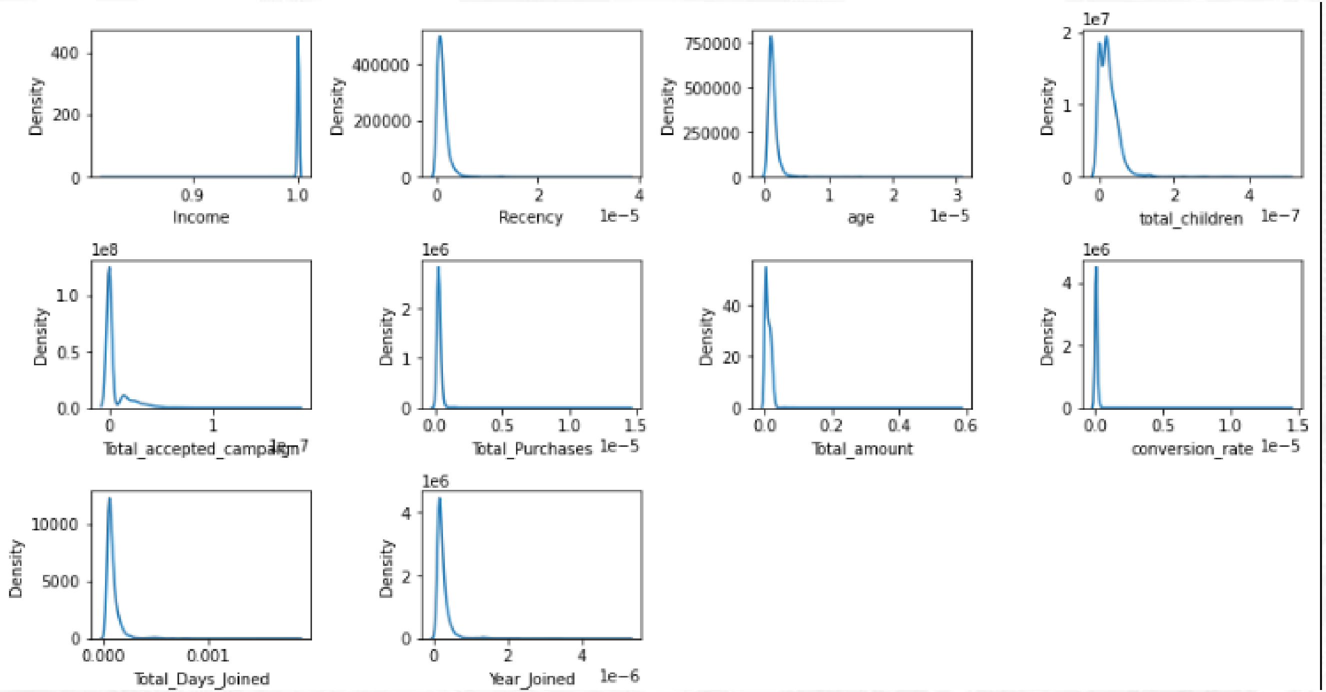


From picture above, we can see range of values between columns has different values, so we need to normalize it so that the result of clustering model will be better.

Click here to see the code

Feature Transformation

For feature transformation, I using normalizer library



Click here to see the code

Rakamin Academy

Modelling K-Means

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

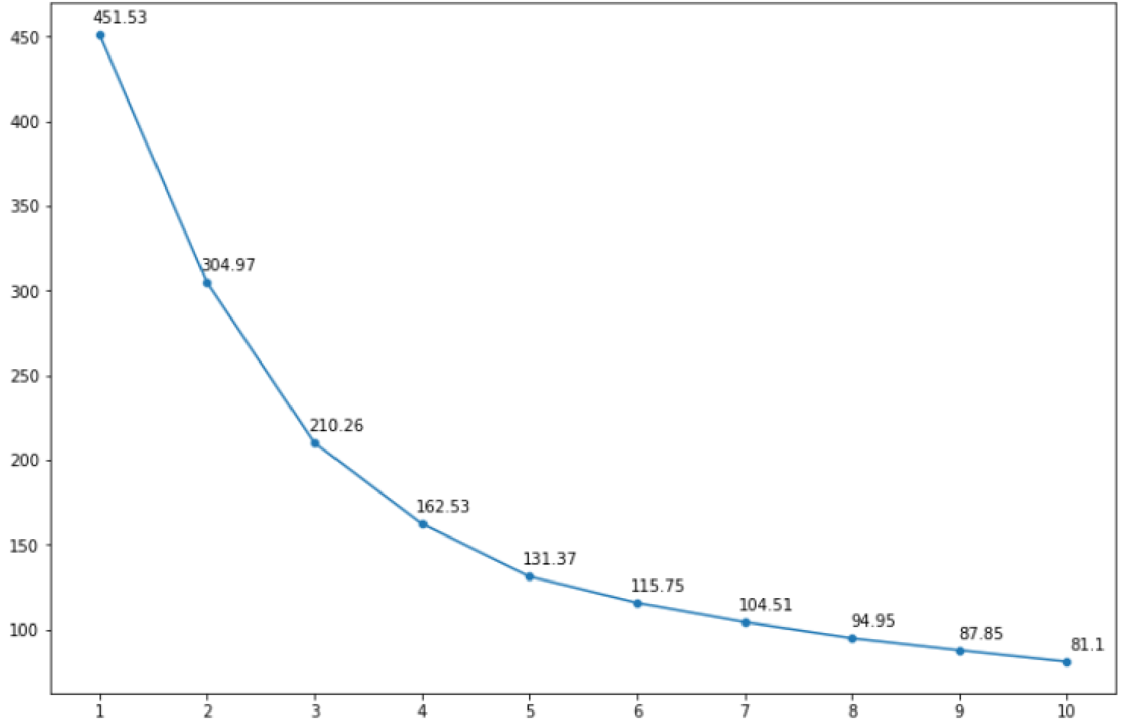
Feature Selection

For segmenting customer, there is a method called RFM Analysis, for you want to know deeply about RFM can read this reference https://www.barilliance.com/rfm-analysis/#:~:text=RFM%20analysis%20is%20a%20data,much%20they've%20spent%20overall.

Name	Meaning	Column that used
Recency	Date of Last of Purchases	Recency
Frequency	Total Number of Orders	Total_Purchases
Monetization	Total order value	Total_amount
Loyalty	Consumer loyalty	Total_accepted_campaign

Click here to see the code

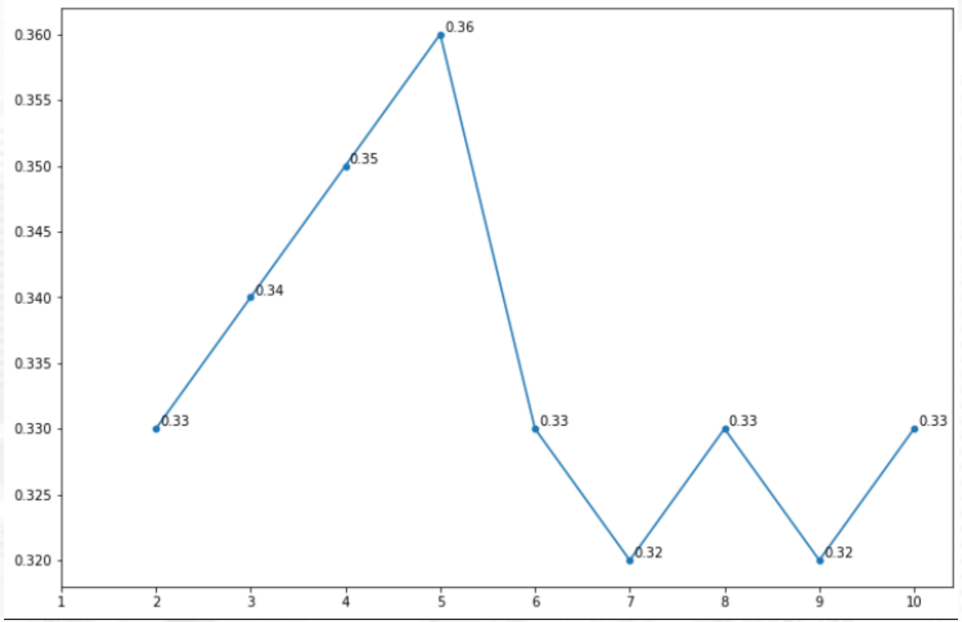
Elbow Score



For picture beside, we can see when n_cluster = 4, the inertia score did'nt change significantly, so we will use n_cluster = 4

[Click here to see the code]

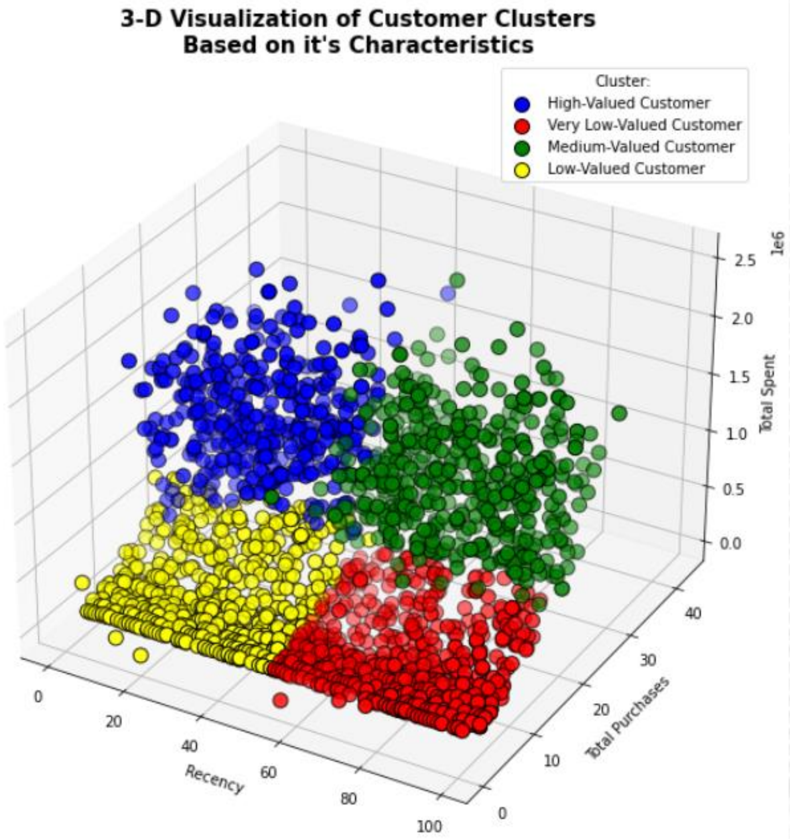
Silhouette score



For picture beside, we can see when n_cluster = 4, it get 0.35 at silhouette score, it's have 0,01 difference with n_cluster = 5 but differ at elbow score we will use n_cluster = 4

[Click here to see the code]

Modelling Result



There are 4 Customer Segments :

1. High Valued Customer
2. Medium Valued Customer
3. Low Valued Customer
4. Very Low Valued Customer

Click here to see the code

Modelling Result

1. High Valued Customer
    - Customers on this group have low average recency (21 days), high average total purchase (22 times) and high average total amount (1,23 Million Rupiah).
    - There 18,44 % of our customer fall into this category.
    - There are 236 customer never accept our campaign, 107 accepted it once, 42 accepted it twice, 20 accepted it three times and 6 accepted it four times.

2. Medium Valued Customer
    - Customers on this group have high average recency (71 days), high average total purchase (22 times) and high average total amount (1,18 Million Rupiah).
    - There 24.09 % of our customer fall into this category.
    - There are 354 customer never accept our campaign, 116 accepted it once, 38 accepted it twice, 24 accepted it three times and 5 accepted it four times.

Click here to see the code

Modelling Result

3. Low Valued Customer
    - Customers on this group have low average recency (23 days), low average total purchase (10 times) and low average total amount (172K Rupiah).
    - There 29.03 % of our customer fall into this category.
    - There are 590 customer never accept our campaign, 54 accepted it once and 3 accepted it twice.
4. Very Low Valued Customer
    - Customers on this group have high average recency (73 days), low average total purchase (10 times) and low average total amount (151K Rupiah).
    - There 28,44 % of our customer fall into this category.
    - There are 587 customer never accept our campaign and 47 accepted it once.

Click here to see the code

Potential Impact

Rakamin Academy

If we keep prioritize on customer cluster and they have similar character like in our data, we still have potential GMV Rp 1.34 Billion with details :

1. High Value Customer : Rp 506 Million
2. Medium Value Customer : Rp 365 Million
3. Low Value Customer : Rp 111 Million
4. Very Low Value Customer : Rp 95 Million

Click here to see the code

Business Recommendation

1. Make a membership (Platinum, Gold, Silver and Bronze) depends on customer cluster, Promote the benefits of being a platinum such as have a more discount to increase our GMV.
2. To Decrease a total customer at Very low and low value customer, we can inform them about our limited discount product and make cheap packages (like buy 1 milk get 1 instant noodle free) since they have a lowest total amount at our shop.
3. To keep our medium and high value customer, we can give them ‘special treatment’ like giving bonuses and gift.

Click here to see the code