Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
Muhammad Iqbal Mudzakky
Muhammad Iqbal Mudzakky
(25) Muhammad Iqbal Mudzakky | LinkedIn

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

Overview

“Sebuah perusahaan dapat berkembang dengan pesat saat mengetahui perilaku customer personality nya, sehingga dapat memberikan layanan serta manfaat lebih baik kepada customers yang berpotensi menjadi loyal customers. Dengan mengolah data historical marketing campaign guna menaikkan performa dan menyasar customers yang tepat agar dapat bertransaksi di platform perusahaan, dari insight data tersebut fokus kita adalah membuat sebuah model prediksi kluster sehingga memudahkan perusahaan dalam membuat keputusan ”

Conversion Rate Analysis Based on Income, Spending and Age

Pada tahap awal, disini saya membuat beberapa feature baru diantaranya :

1. ‘convertion_rate’ = ratio antara jumlah customer yang respon pada feature ‘Response’ dengan jumlah customer keseluruhan per kategori
2. ‘total_spending’ = jumlah dari feature ‘MntCoke’, ‘MntFruits’, ‘MntMeatProducts’, ‘MntFishProducts’, ‘MntSweetProducts’, ‘MntGoldProds’
3. ‘total_spending_category’ = kategori berdasarkan jumlah transaksi yang dikeluarkan oleh customer (‘total_spending’), dimana :
    - ‘low’ : jika jumlah transaksi < 395.000
    - ‘mid’ : jika jumlah transaksi < 1.040.000 tetapi >= 395.000
    - ‘high’ : jika jumlah transaksi >= 1.040.000

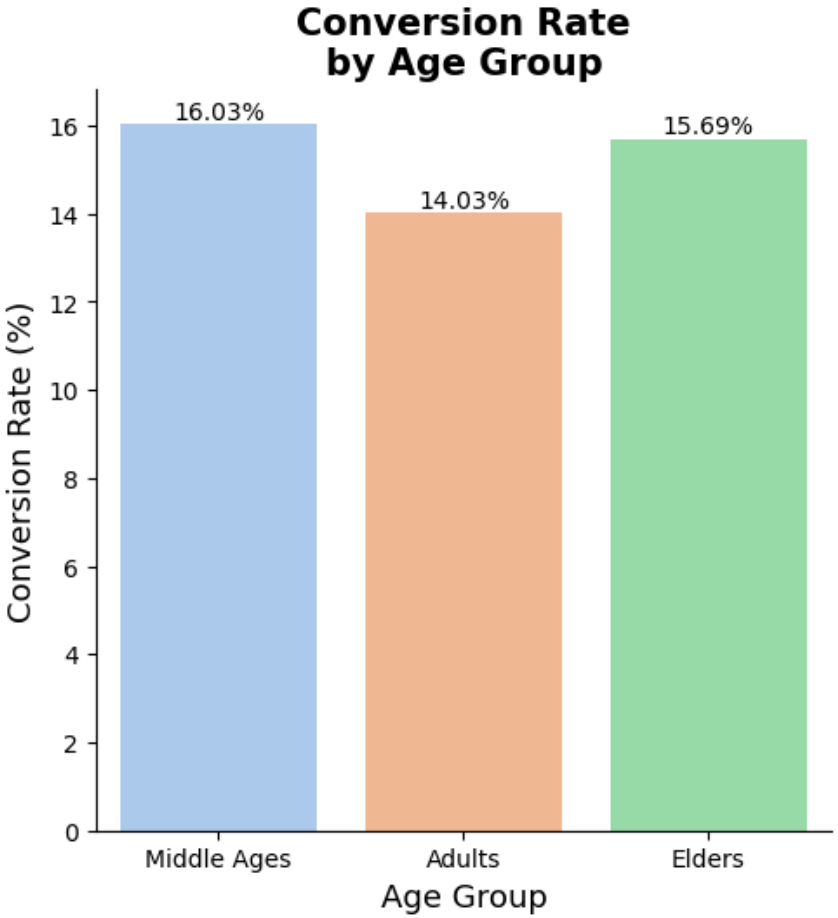
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

4. ‘income_category’ = membuat kategori berdasarkan besaran income dimana :
   - ‘low’ : jika nilai income < 51.000.000
   - ‘mid’ : jika nilai income >= 51.000.000 tetapi < 68.000.000
   - ‘high’ : jika nilai income >= 68.000.000
5. ‘years’ = merupakan extract value dari feature ‘Dt_Customer’ dimana value tersebut di split dan hanya diambil tahun nya saja menggunakan function lamba x.split(‘-’)[2]
6. ‘age’ = feature yang merepresentasikan umur (feature ‘years’ - feature ‘Year_Birth’)
7. ‘age_category’ = membuat kategori berdasarkan umur dimana :
   - ‘Middle Age’ : rentan umur < 40
   - ‘Adults’ : rentan umur < 60 tetapi >= 40
   - ‘Elders’ : rentan umur >= 60

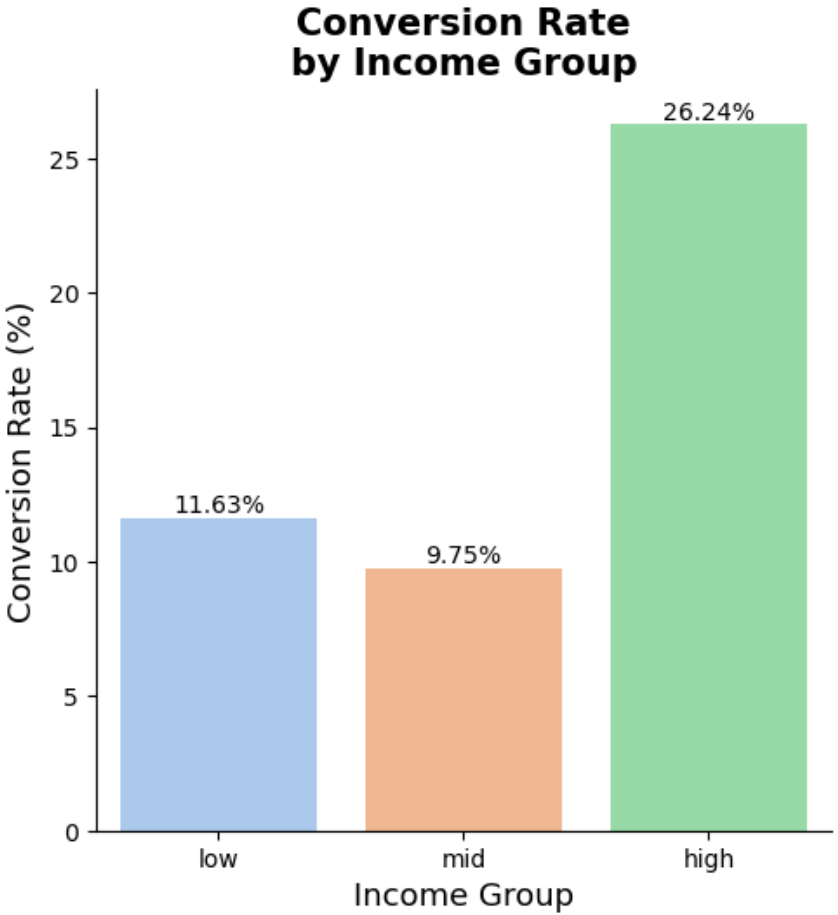
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Age



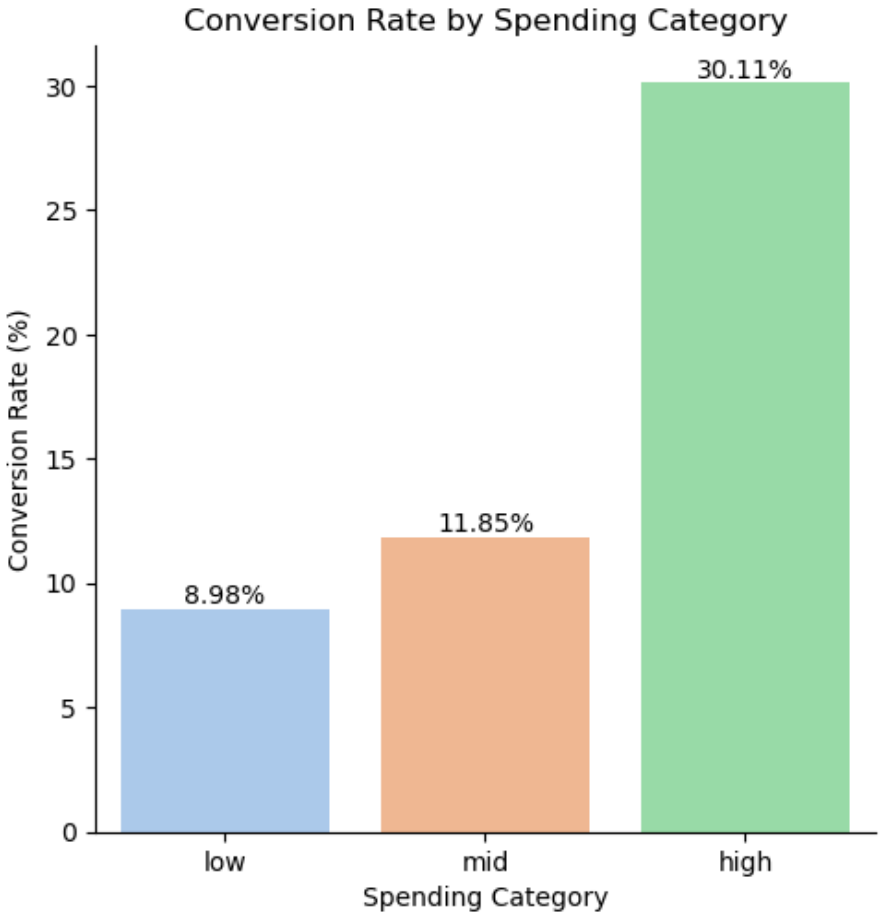
- Jika dilihat berdasarkan grafik disamping, perbedaan umur dari customer tidak terlalu berbeda secara signifikan untuk nilai conversion rate nya, dimana nilai dari conversion rate itu sendiri berkisar 14% - 16%.
- Berdasarkan hasil Analisa, untuk meningkatkan nilai conversion rate sebaiknya tidak terlalu terpaku pada rentan umur customer, karena jika dilihat dari gambar disamping, perbedaan umur customer memiliki nilai conversion rate yang tidak terlalu berbeda jauh.

Conversion Rate Analysis Based on Income



- Jika dilihat berdasarkan grafik disamping, perbedaan income dari customer mempengaruhi nilai conversion rate nya, dimana untuk nilai conversion rate tertinggi ada pada kelompok high income sebesar ~26%.
- Berdasarkan hasil Analisa, untuk meningkatkan nilai conversion rate ada baiknya jika tim marketing memprioritaskan target customer yang masuk ke dalam kategori high income, dimana nilai income customer pada kategori ini berkisar > 68.000.000

Conversion Rate Analysis Based on Income



- Jika dilihat berdasarkan grafik disamping, perbedaan jumlah transaksi dari customer mempengaruhi nilai conversion rate nya, dimana untuk nilai conversion rate tertinggi ada pada kategori high sebesar ~30%.
- Berdasarkan hasil Analisa, untuk meningkatkan nilai conversion rate ada baiknya jika tim marketing memprioritaskan target campaign customer yang masuk ke dalam kategori high, dimana jumlah transaksi customer pada kategori ini berkisar > 1.040.000

Data Cleaning & Preprocessing

- Tahap pertama yang dilakukan adalah mengecek *missing value* dengan script :

```
df.isna().sum()
```

Kemudian didapatkan hasil bahwa terdapat *missing value* pada feature Income sebanyak 24 data dari keseluruhan data 2240 data (~1%).
Karena nilai missing value ini <5%, disini saya drop menggunakan script :

```
df = df.dropna()
```

- Tahap kedua, saya melakukan pengecekan terhadap *duplicated value* dengan script

```
df.drop(columns='ID',inplace=True)
if df.duplicated().sum()==0: print('tidak ada data duplicated')
else : print(f'terdapat data duplicated sebanyak
:',df.duplicated().sum())
```

Kolom 'ID' di drop karena merupakan kolom unique, dimana jika feature ini diikutsertakan dalam modeling akan mengganggu performance dari model machine learning.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Didapat pada data terdapat *duplicated value* sebanyak 183, selanjutnya di drop menggunakan script :

```
df = df.drop_duplicates()
```

- Tahap ketiga, melakukan StandardScaler. Transformasi ini mengubah mean dari sebaran data menjadi 0 dengan standar deviasi = 1. Transformasi ini biasa dilakukan untuk data yang memiliki distribusi sebaran data positively skewed (mean > median). Adapun script yang digunakan adalah :

```
from sklearn.preprocessing import StandardScaler

ss = StandardScaler()
col = ['Income', 'MntCoke', 'MntFruits', 'MntMeatProducts',
'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'dt_years']

for i in col :
    df[i] = ss.fit_transform(df[i].values.reshape(-1,1))
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature yang dilakukan transformasi hanya feature yang memiliki value numeric berbeda jauh dengan feature lainnya. Hal ini dilakukan supaya pada saat melakukan clustering nanti, jarak antar cluster tidak timpang dengan value dari feature2 yang lain, sehingga memudahkan untuk clustering dan hasilnya lebih optimal.

- Tahap keempat melakukan *feature encoding* dimana ini membuat *data type object* menjadi *integer*. Sebagaiman diketahui, karena clustering ini merupakan algorithma yang mengelompokan value berdasarkan jarak, maka data yang digunakan juga harus dalam bentuk *integer* bukan *string*. Dalam tahap ini script yang digunakan :

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
columns = df.select_dtypes(include='object').columns
for i in columns:
    df[i] = le.fit_transform(df[i])
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Value yang diubah pada tahap ini ada 2 feature, yang pertama 'Education' dimana masing2 value :

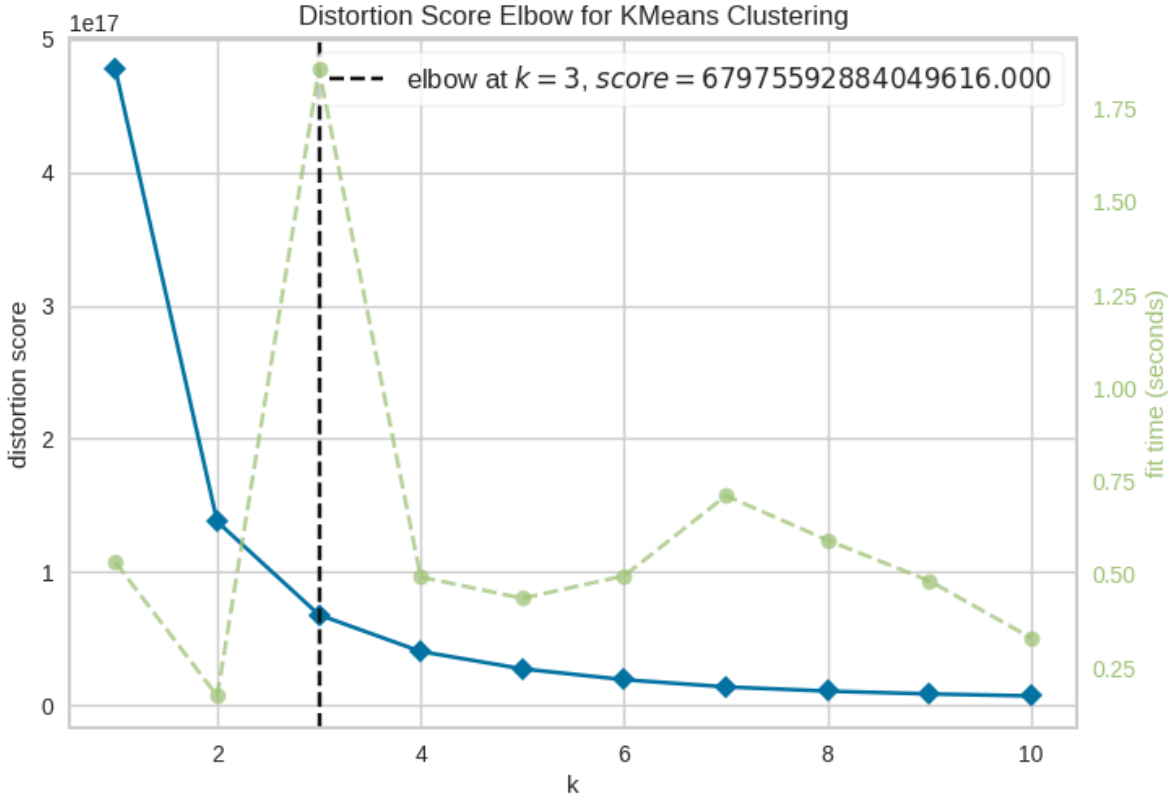
- S1 = 1
- S3 = 3
- S2 = 2
- D3 = 0
- SMA = 4

Yang kedua, feature 'Marital_Status' dimana untuk masing2 value:

- Menikah = 5
- Bertunangan = 0
- Lajang = 4
- Cerai = 1
- Janda = 3
- Duda = 2

Untuk selengkapnya, dapat melihat jupyter notebook disini

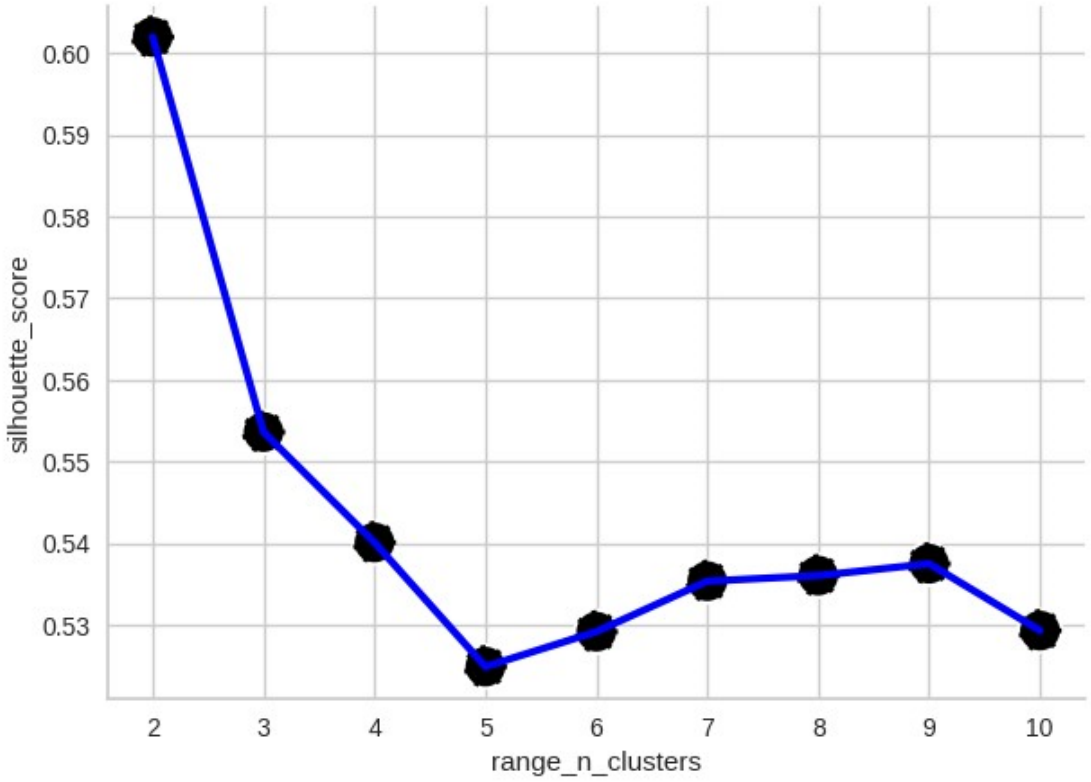
Data Modeling



Pada plot disamping, berdasarkan elbow method yang digunakan, penurunan nilai distortion score (nilai inertia = ukuran yang menggambarkan seberapa jauh data poin dalam 1 kluster terhadap pusat klusternya) signifikan terjadi sampai titik k=3, dimana setelah nya untuk nilai distortion score penurunannya tidak mengalami signifikan. Hal ini menandakan titik siku dari plot tersebut terdapat pada k=3, sehingga jika menggukan elbow method rekomendasi kluster dari dataset tersebut dibagi menjadi 3 kluster

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling



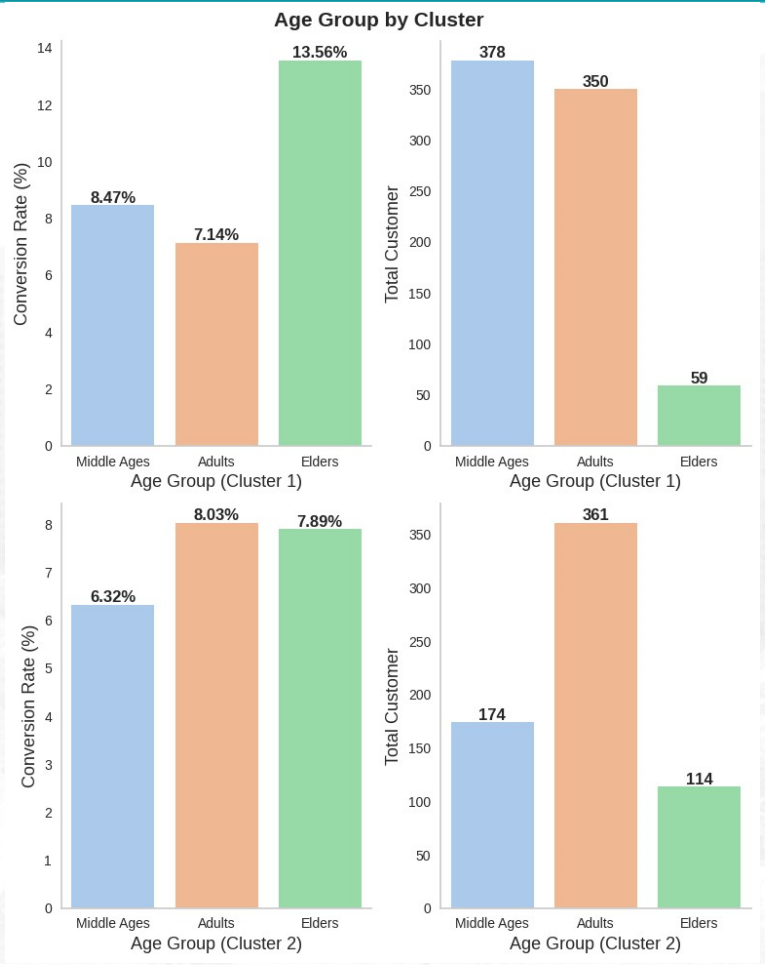
Plot disamping menggambarkan perbandingan antara metrics silhouette score dengan jumlah kluster. Silhouette Score mengukur seberapa dekat setiap data poin dengan kluster tempat mereka berada dibandingkan dengan kluster lainnya. Rentang nilai Silhouette Score adalah -1 hingga 1, di mana nilai positif menunjukkan bahwa objek berada di kluster yang tepat, sedangkan nilai negatif menunjukkan bahwa objek mungkin ditempatkan di kluster yang salah. Pada plot silhoutter score tertinggi ada pada range kluster 2 dengan nilai score > 0.6.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling

Penggunaan algorithma K-Means untuk evaluasi score dan pengambilan keputusan jumlah kluster bukan hanya bisa dengan menggunakan elbow method atau silhouette score saja, tetapi masih ada metrics lain seperti SSE (Sum of Squared Errors), Dunn Index dan lainnya. Oleh karena itu, benar atau salah nya penentuan jumlah kluster dapat ditinjau kembali dari karakteristik data tersebut. Dalam kasus ini, jumlah k=2 lebih bisa diimplementasikan disbanding k=3. Karena untuk k=2, kita dapat dengan jelas untuk melakukan segmentasi customer berdasarkan feature yang ingin kita gunakan, dalam hal ini feature kategori income dan total spending merupakan feature yang kita gunakan dalam menganalisa target market kedepannya

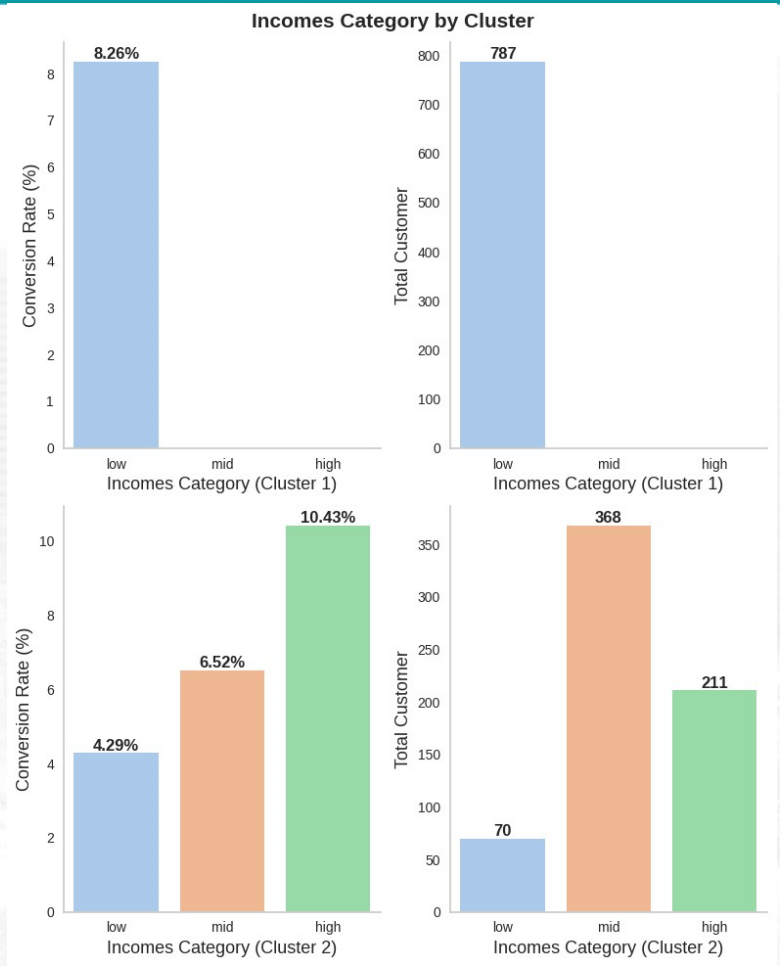
Untuk selengkapnya, dapat melihat jupyter notebook disini



Age Category

Berdasarkan segmentasi kategori umur untuk setiap cluster, untuk **cluster 1** jumlah kategori umur terbanyak customer adalah **Middle Ages** (selisih 20 dengan Adults) dimana nilai conversion rate sebesar **8.47%**, sedangkan untuk **cluster 2** jumlah kategori umur terbanyak adalah **Adults** dengan nilai conversion rate sebesar **8.03%**.

Untuk selengkapnya, dapat melihat jupyter notebook disini

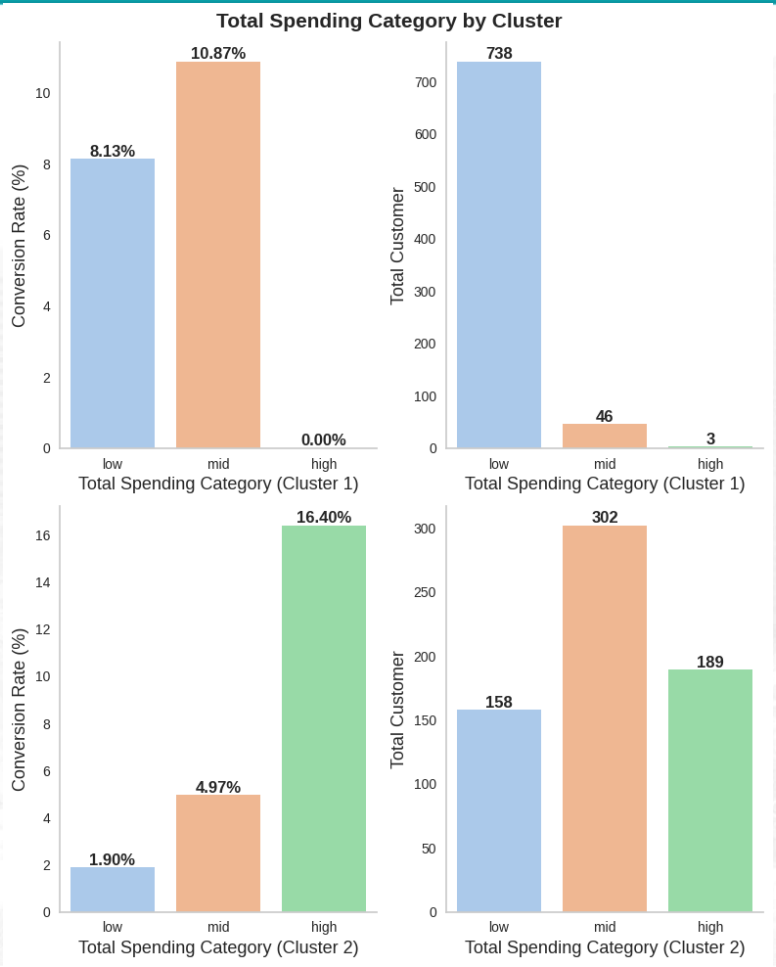


Income Category

Berdasarkan segmentasi kategori pendapatan untuk setiap cluster, untuk **cluster 1** jumlah kategori pendapatan terbanyak customer adalah **low** dimana nilai conversion rate sebesar **8.26%**, sedangkan untuk **cluster 2** jumlah kategori pendapatan terbanyak adalah **mid** dengan nilai conversion rate sebesar **8.03%**.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting



Total Spending Category

Berdasarkan segmentasi kategori total pengeluaran untuk setiap cluster, untuk **cluster 1** jumlah kategori pengeluaran terbanyak customer adalah **low** dimana nilai conversion rate sebesar **8.13%**, sedangkan untuk **cluster 2** jumlah kategori pendapatan terbanyak adalah **mid** dengan nilai conversion rate sebesar **4.97%**.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conclusion

1. **Cluster 1**

Memiliki karakteristik customer Middle Ages Category (**range umur < 40 tahun**), Low Income Category (**pendapatan < 51.000.000**) dan Low Total Spending Category (**total belanja < 395.000**).

2. **Cluster 2**

Memiliki karakteristik customer Adults Ages Category (**range umur 40 – 59 tahun**), Mid Income Category (**pendapatan 51.000.000 – 67.999.999**) dan Mid Total Spending Category (**total belanja < 395.000 – 1.039.999**).

Business Recomendation

Dengan mengetahui karakteristik masing2 cluster ini, dapat dilakukan penyesuaian strategi pemasaran untuk memenuhi kebutuhan dan preferensi setiap kelompok pelanggan. Sebagai contoh, untuk Cluster 2 terdiri dari pelanggan dengan pendapatan tinggi dan total pengeluaran yang besar, tim marketing dapat mempertimbangkan untuk menawarkan produk atau layanan premium kepada mereka. Sementari itu, untuk Cluster 1 dengan karakteristik yang berlawanan mungkin lebih responsif terhadap penawaran diskon atau promosi khusus

Untuk selengkapnya, dapat melihat jupyter notebook disini