Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
Ali Imran Nasution
imransutee@gmail.com
www.linkedin.com/in/imrannasution

"Fresh graduate with Bachelor of Informatics Engineering from Malikussaleh University with a focus on Data Mining and Decision Support Systems. Currently, I have just completed a data science bootcamp at Rakamin Academy with various projects. Currently interested in starting a career as a new data scientist."

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

Feature Engineering

- ***'Age' :*** Mengurangkan tahun saat ini di dataset dengan tahun lahir dari kolom *'year_birth'*
- ***'Children' :*** Menjumlahkan kolom *'kidhome'* dengan *'teenhome'*
- ***'Is_parent'*** *:* Bernilai ***True*** Apabila customer memiliki *'children'*
- ***'Age_Group'*** *:* Mengkategorikan kolom *'Age'* menjadi 5 kategori yaitu : 18-24, 25-34, 35-44, 45-54, 55-64, dan Olders
- '**Total_Spend**' : Menjumlahkan kolom *'MntCoke*', *'MntFruits*', *'MntMeatProducts*', *'MntFishProducts*, *'MntSweetProducts*', dan *'MntGoldProds*'
- '**Total_Campaign**' : Menjumlahkan kolom *'AcceptedCmp1*', *'AcceptedCmp2*', *'AcceptedCmp3*', *'AcceptedCmp4*', dan *'AcceptedCmp6*'
- '**Total_Transaction**' : Menjumlahkan kolom 'NumDealsPurchases', 'NumCatalogPurchases', 'NumStorePurchases', dan 'NumWebVisitsMonth'
- '**Member_Duration**' : Mengurangkan tahun saat ini di dataset dengan kolom '*Dt_customer*'
- '**Coversion_Rate**' : Membagi kolom #'*response*' / #'*NumWebVisitsMonth'*

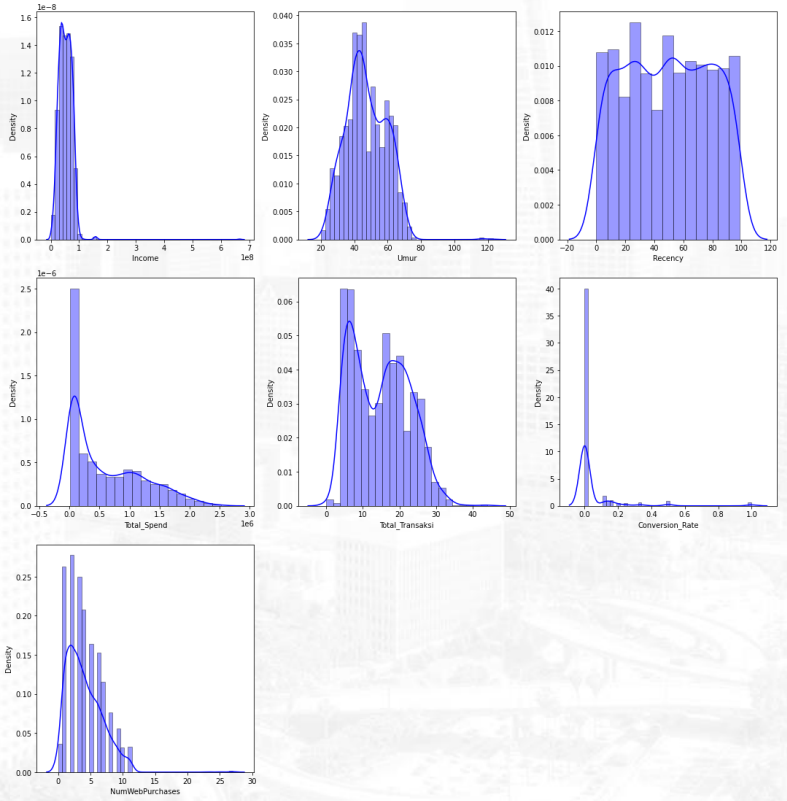
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

Exploratory Data Analysis

Univariate Analysis

- Terlihat kolom income, Total_Spend, Total_Campaign, Conversion_Rate, NumWebPurchases mengalami skewness
- Nilai dari masing-masing kategori sangat besar sehingga perlu di scaling
- Kebanyakan customer memiliki income dikisaran 6.000.000 - 8.000.000
- Customer berada dalam usia antara 20 hingga 80, 120 kemungkinan outliers
- Recency dari customer diantara 10 - 90
- Total transaksi terbanyak antara 4 – 7 kali
- Total Spend yang dihabiskan customer dikisaran 20000 - 50000
- Terlihat conversion rate sangat rendah
- Customer mengunjungi web bulanan mayoritas di kisaran 1 - 4



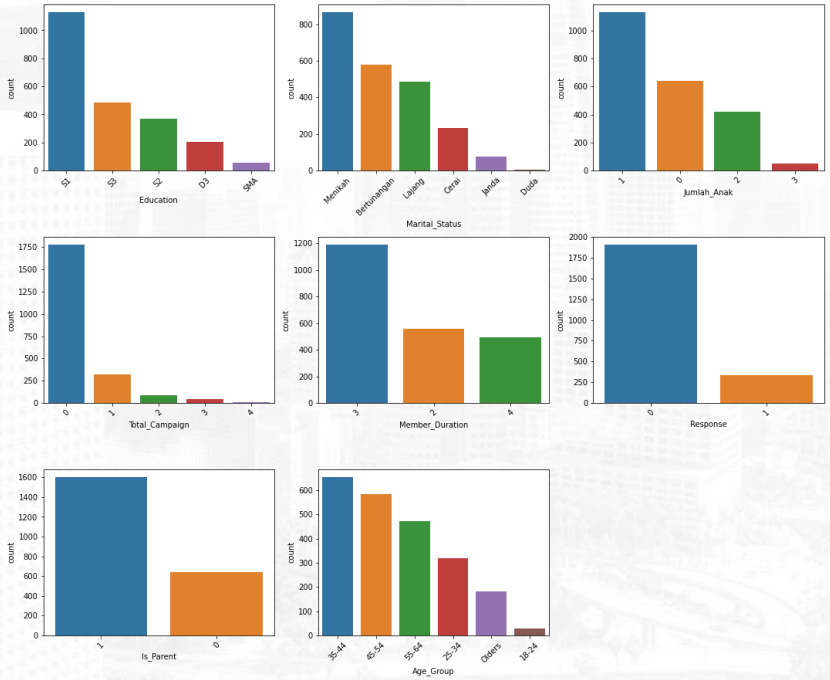
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

Exploratory Data Analysis

Univariate Analysis

- Customer didominasi oleh latar pendidikan S1
- Customer didominasi oleh hubungan yang memiliki ‘ikatan’
- Terlihat bahwa kebanyakan customer menolak response dan campaign.
- Sebagian besar customer adalah orang tua yang memiliki satu anak
- Customer didominasi oleh umur >35an keatas
- Customer berlangganan paling lama 4 tahun akan tetapi di kebanyakan 3 tahun.



Untuk selengkapnya, dapat melihat jupyter notebook disini

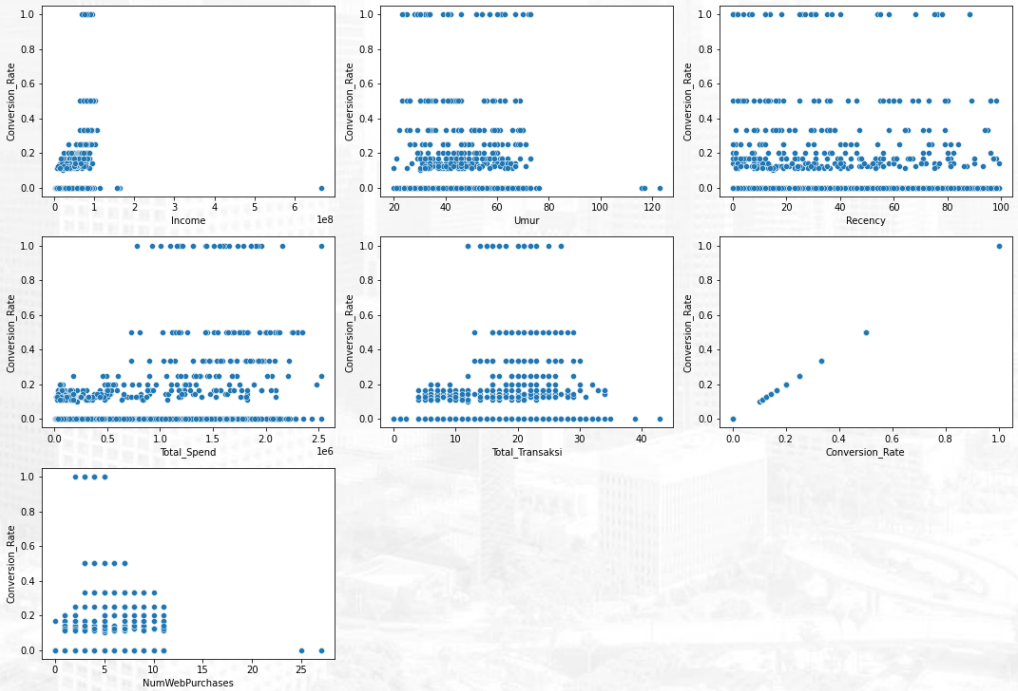
Conversion Rate Analysis Based on Income, Spending and Age

Exploratory Data Analysis

Bivariate Analysis

- Tidak ada kolom yang berkolerasi baik dengan conversion_rate / weak correlation

- ada hubungan yang memiliki korelasi antara
    - income - total_spend
    - income - total_transaksi
    - income - NumWebPurchases
    - NumWebPurchases - Total Transaksi
    - total_spend - total_transaksi



Untuk selengkapnya, dapat melihat jupyter notebook disini

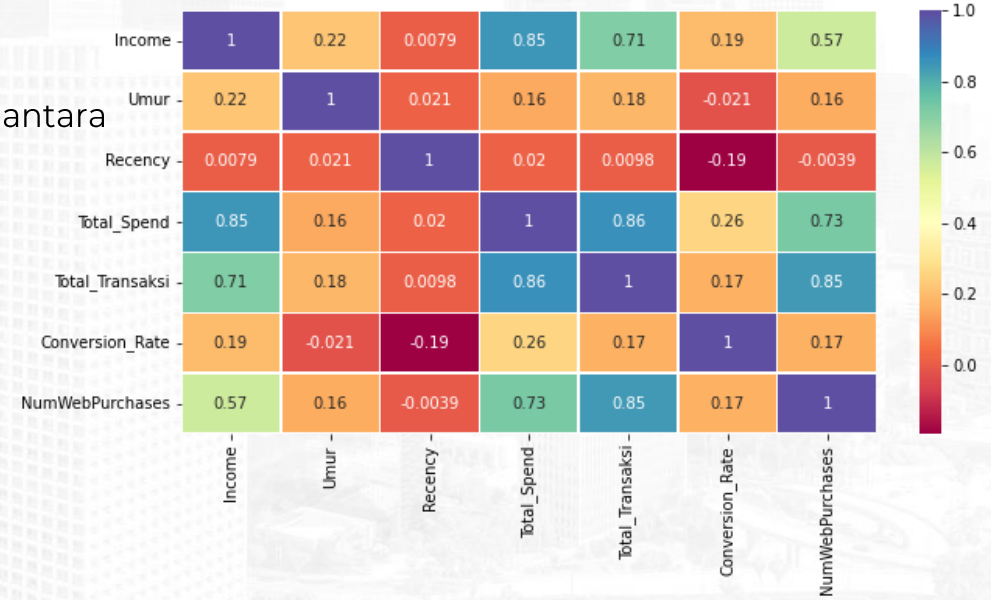
Conversion Rate Analysis Based on Income, Spending and Age

Exploratory Data Analysis

Multivariate Analysis

- Terlihat kolom yang memilki korelasi tinggi antara lain :
  - Income - Total Spend,
  - Income - Total Transaksi,
  - Total_Spend - NumWebPurchases,
  - Total_Spend - Total_Transaksi
  - Total_Transaksi – NumWebPurchases

- Kolom yang memiliki korelasi yang moderate :
  - Income - NumWebPurchases



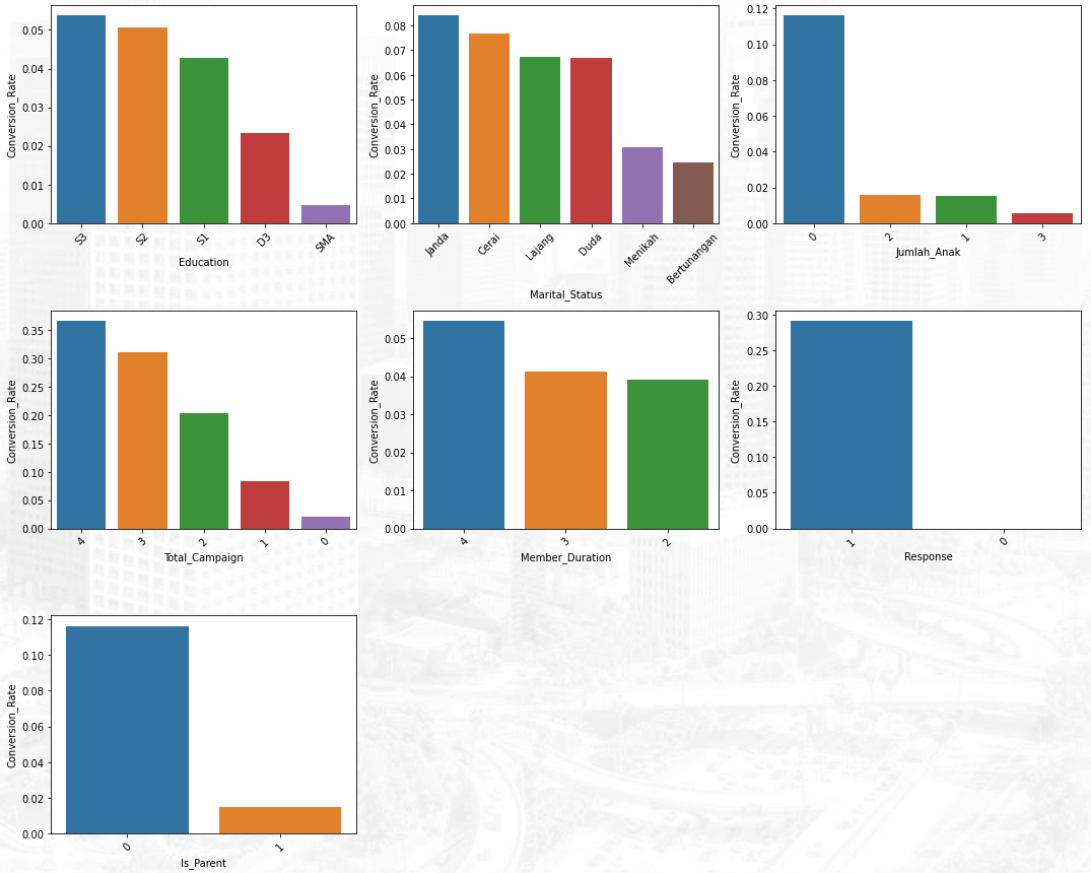
Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age

Bussiness Insight

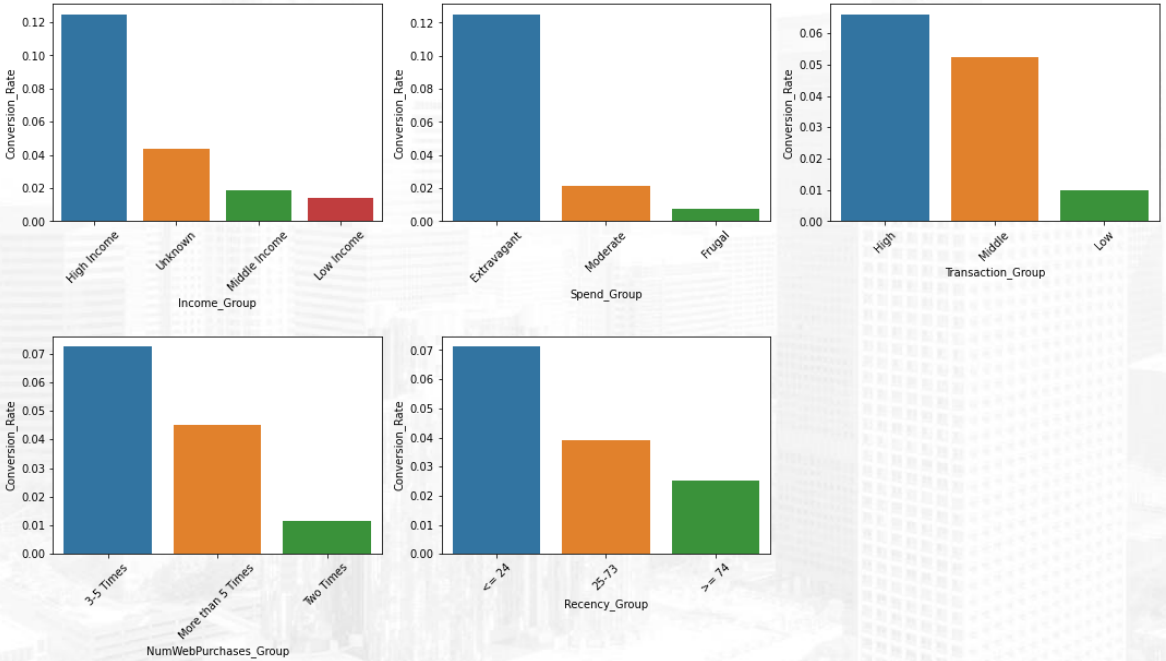
Karakteristik Customer yang memiliki conversion rate yang tinggi :

- Terlihat bahwa mayoritas customer berstatus marital 'yang mengikat' namun secara CR justru 'hubungan yang tidak mengikat yang lebih tinggi'
- Secara usia mayoritas pengguna berumur 34 - 45, akan tetapi umur yang 18 - 34 yang memiliki CR yang lebih tinggi/ merespon campaign
- Total Campaign 4 atau 3
- Tidak memiliki anak/bukan orang tua
- Pendidikan Tinggi yaitu dari S1 - S3



Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income, Spending and Age



- High Income yaitu lebih besar dari 68.522.000
- Total Spend lebih besar dari 1.045.500 atau Extravagant Customer
- Total Transaksi lebih dari 4 kali
- Mengunjungi Halaman Web antara 3-5 kali
- Nilai Recency dibawah 25
- Sangat masuk akal apabila yang memiliki income tinggi akan cenderung merespon campaign di karenakan customer ini biasanya memiliki uang yang lebih dengan dibuktikan total spend yang lebih tinggi dengan total transaksi yang terbilang sering, hal ini bisa dilihat dari jumlah kunjungan ke halaman web.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning

Missing Value

Terdapat 24 missing value pada kolom Income. Adapun threatment yang dibelikan
Adalah melakukan imputasi median kolom Income pada missing value tersebut.

```
# imputasi median
dfx['Income'].fillna(dfx['Income'].median(), inplace=True)
```

```
# check null
df.isnull().sum()
```

```
ID                   0
Year_Birth           0
Education            0
Marital_Status       0
Income              24
Kidhome              0
Teenhome             0
Dt_Customer          0
```

Duplicate Value

Tidak terdapat duplikasi record pada data.

```
#Check duplicates value
df.duplicated().sum()
```

```
0
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

Feature Selection

Adapun feature yang dipilih adalah hasil dari feature engineering dengan teknik mengurangi dimensi seperti lebih memilih total campaign dibanding AcceptedCmp1, AcceptedCmp2 dst. Adapun feature yang dipilih antara lain : *Income* , *Recency*, *NumWebPurchases*, *Response*, *Jumlah_Anak*, *Is_Parent*, *Total_Spend*, *Total_Campaign*, *Total_Transaksi*, *Member_Duration*, *Conversion_Rate*

Handling Outliers

Adapun handling outliers yang dilakukan menggunakan Teknik Z-Score. Sebelum dilakukan penangan outliers data berjumlah 2240, setelah dilakukan penangan data berkurang menjadi 2222.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

Feature Encoding

Terdapat dua kolom kategorikal yang belum dilakukan encoding, yaitu Education dan Marital_Status. Data pada kolom education adalah ordinal categorical sehingga perlu dilakukan encoding dengan cara mapping sesuai tingkat hirarkinya. Sedangkan Marital_Status adalah nominal categorical, Teknik yang dilakukan adalah One Hot Encoder.

Feature Standarization

Nilai di beberapa feature memiliki perbedaan rentang nilai yang sangat besar sehingga perlu distandarisasi agar normal. Pada kolom yang tipe data numerical digunakan Teknik 'StandardScaler' dan 'MinMax' untuk beberapa kolom kategorikal yang memiliki nilai unik lebih dari 2 seperti Marital_Status, Jumlah_Anak, Education, Member_Duration.

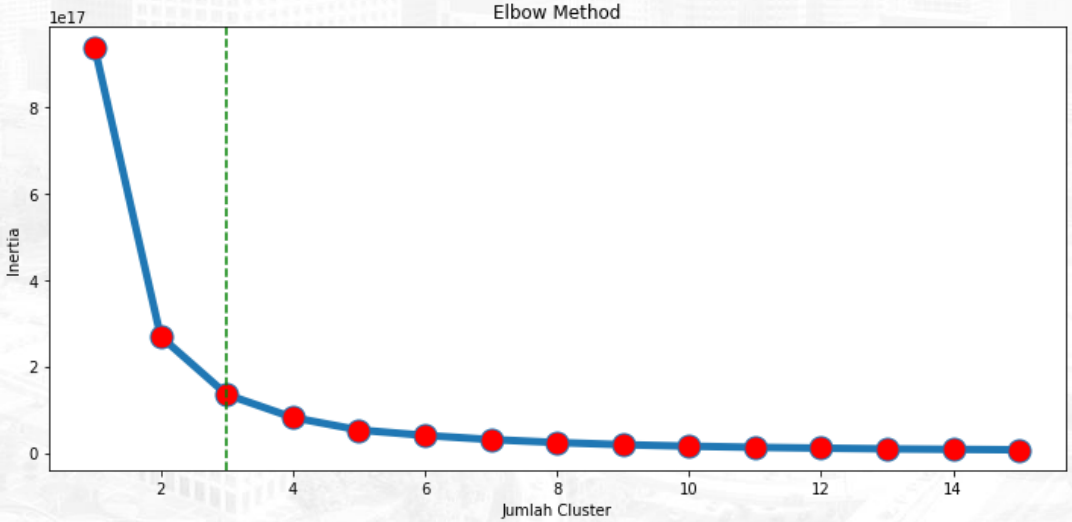
Untuk selengkapnya, dapat melihat jupyter notebook disini

K-Means Algorithm

Elbow Method

Elbow method adalah Teknik yang digunakan untuk mencari nilai k atau jumlah klaster yang paling optimal pada algoritma k-means dengan melihat *siku*.

Terlihat grafik di samping seperti *siku*,
Adapun siku terletak pada k=3,
Dengan kata lain nilai k yang paling
Optimal adalah 3.



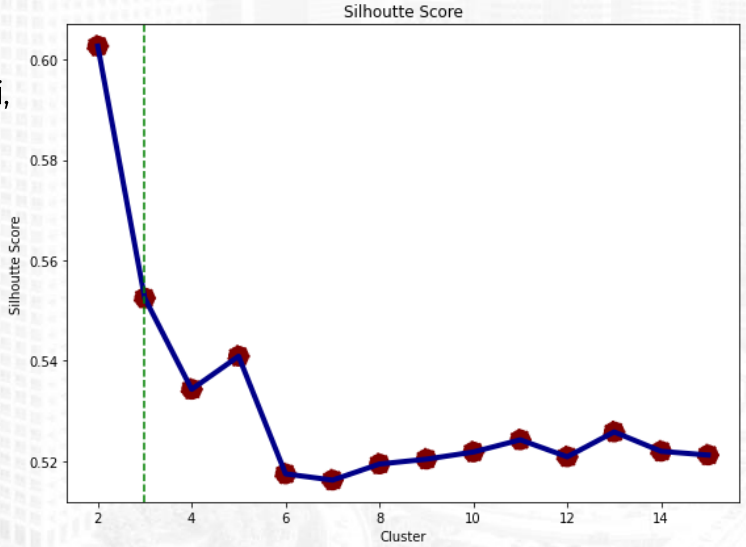
Untuk selengkapnya, dapat melihat jupyter notebook disini

K-Means Algorithm

Evaluate with Silhoutte Score

Silhoutte score adalah ukuran seberapa mirip suatu titik data di dalam klaster (kohesi) dibandingkan dengan klaster lain (pemisahan).

Pada silhouette score diambil nilai rata-rata tertinggi,
Terlihat bahwa k=2 adalah nilai tertinggi,
Tetapi apabila menggunakan dua klaster rasanya terlalu sedikit, oleh karena itu opsi terbaik adalah 3 klister.



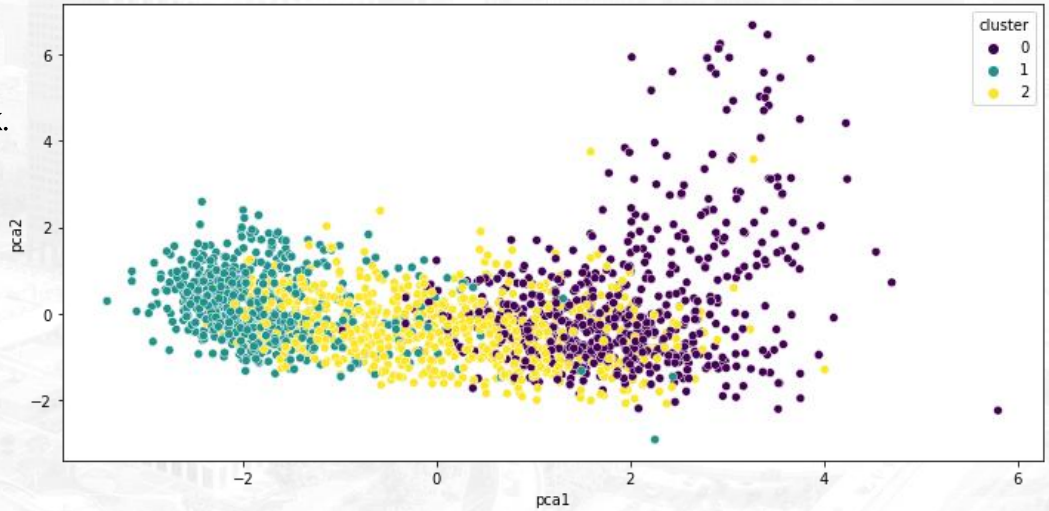
Untuk selengkapnya, dapat melihat jupyter notebook disini

K-Means Algorithm

Cluster Visualization with PCA

Setelah selesai melakukan klastering dengan nilai k=3, kita dapat melihat sebaran data menggunakan Teknik PCA untuk mengetahui seberapa baik pemisahan antar klaster.

Terlihat pada grafik di samping, bahwa Pemisahan antar klister sudah cukup baik.



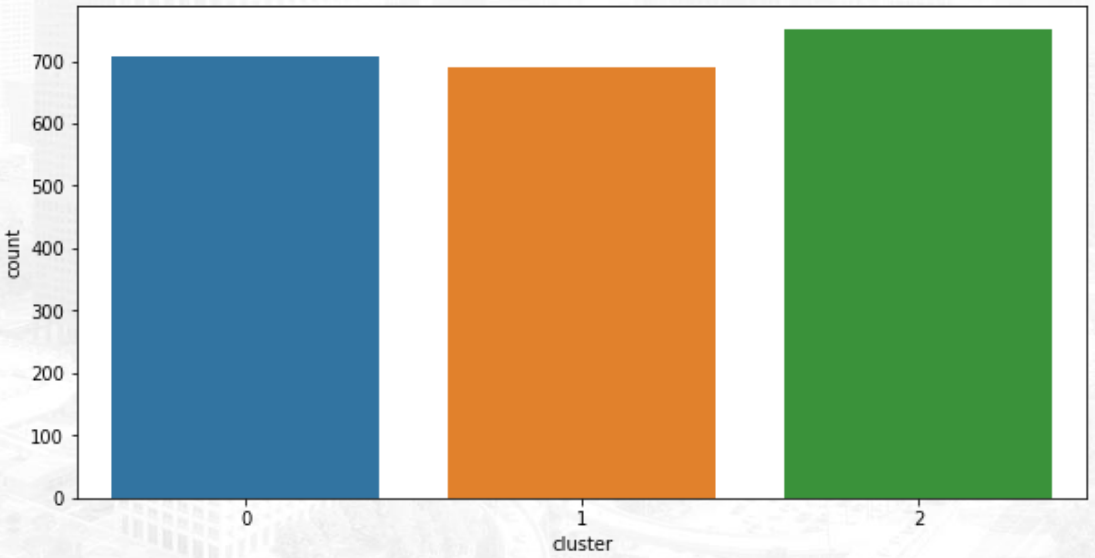
Untuk selengkapnya, dapat melihat jupyter notebook disini

Num of Customer by Clusters.

Hasil dari klaster sebelumnya dapat kita lihat sebaran datanya sebagai berikut :

Terlihat bahwa perbedaan total csutomer antar claster hampir sama.

Dengan kata lain bahwa proporsi antar Klaster hampir seimbang.



Untuk selengkapnya, dapat melihat jupyter notebook disini

Cluster Insight

Untuk mengetahui karakteristik dari masing-masing klaster, kita dapat meng-agregasi tabel untuk mencari nilai rata-ratanya sebagai berikut :

Cluster	Income	Umur	Recency	Total_Spend	Total_Transaksi	Conversion_Rate	NumWebPurchases
0	27948220.65	42.77	48.32	88927.86	7.74	0.02	2.06
1	73715852.39	49.49	50.13	1209480.46	21.00	0.04	5.58
2	50580615.18	49.27	49.85	429470.04	15.30	0.02	4.51

Untuk selengkapnya, dapat melihat jupyter notebook disini

Cluster Insight

Untuk mengetahui karakteristik dari masing-masing klaster, kita dapat meng-agregasi tabel untuk mencari nilai rata-ratanya sebagai berikut :

Cluster	Income	Umur	Recency	Total_Spend	Total_Transaksi	Conversion_Rate	NumWebPurchases
0	27948220.65	42.77	48.32	88927.86	7.74	0.02	2.06
1	73715852.39	49.49	50.13	1209480.46	21.00	0.04	5.58
2	50580615.18	49.27	49.85	429470.04	15.30	0.02	4.51

Untuk selengkapnya, dapat melihat jupyter notebook disini

Cluster Insight

Apabila kita lihat dari rata-rata pada masing-masing klaster, beberapa kolom perbedaannya sangat mencolok yaitu:

Income
Cluster 0 : Low Income
Cluster 1 : High Income
Cluster 2 : Middle Income

Total_Spend
Cluster 0 : Frugal
Cluster 1 : Extravagant
Cluster 2 : Moderate

Total_Transaksi
Cluster 0 : Low Transaction
Cluster 1 : High Transaction
Cluster 2 : Middle Transaction

Dari kolom tersebut dapat disimpulkan bahwa :

- Cluster 0 : Low Income, Frugal, Low Transaction => Iron
- Cluster 1 : High Income, Extravagant, High Transcation => Platinum
- Cluster 2 : Middle Income, Moderate, Middle Transaction => Gold

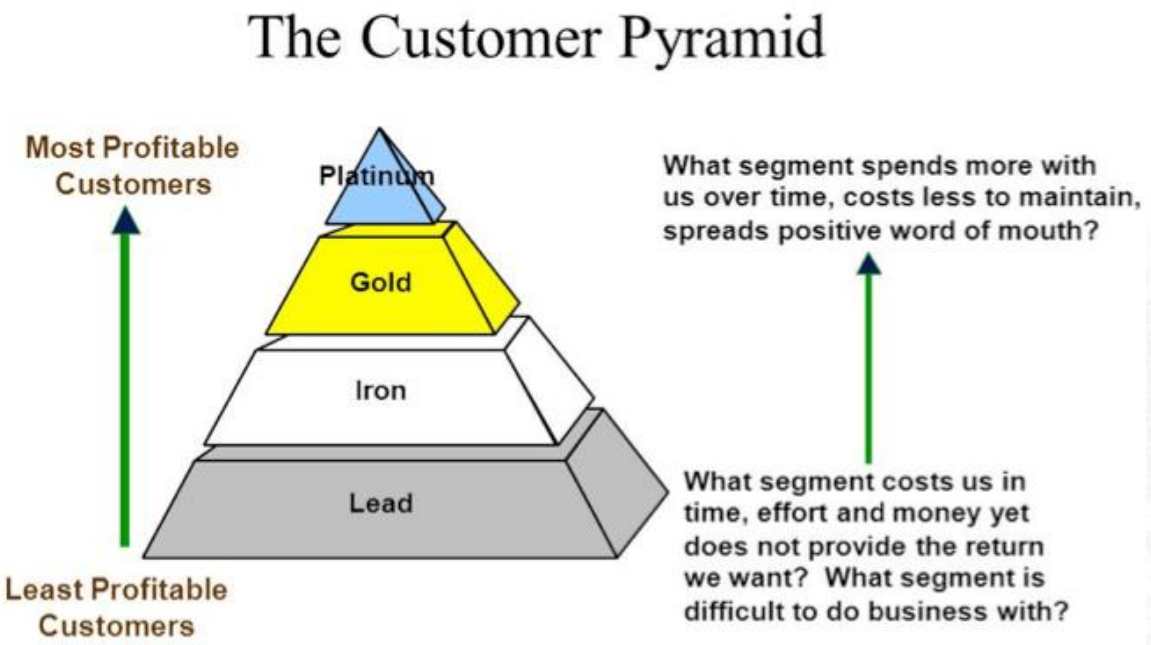
Untuk selengkapnya, dapat melihat jupyter notebook disini

Potential Impact

Sesuai dari piramida, maka Customer yang memiliki Potential impact yang paling Tinggi adalah Platinum dengan Rata-rata total spend sebesar 1.209.480.

Adapun Saran Bisnis yang bisa dikembangkan :

1. Program Loyalti/Privilaged
2. Model Retensi



Untuk selengkapnya, dapat melihat jupyter notebook disini

Bussiness Recomendation

1. Platinum Customers as Most Profitable Customer
Program loyalitas, program frekuensi, dan program insentif lainnya adalah strategi umum untuk mempertahankan customer inti ini.

2. Gold and Iron Customers
Menargetkan customer ini dengan program insentif atau penghargaan serupa, dan melakukan hal-hal untuk memperkuat hubungan dan mendorong mereka ke piramida pelanggan.

Untuk selengkapnya, dapat melihat jupyter notebook disini