Predict Customer Personality to boost marketing campaign by using Machine Learning

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

Rakamin Academy

Created by:
Muchammad Malik
muchammad.malik@gmail.com
https://www.linkedin.com/in/muchammad-malik/

“Currently working as business and system development at GESITS, the leading EV brand in Indonesia. Having a bachelor degree from engineering physics , I possessed balanced skill in engineering and management discipline.

As a data science and business analyst enthusiast, I developed skillset in Business Accumen, SQL, Pyhton, Tableau, and machine learning. I have spent 2 years to learn and maintain these skill by taking several bootcamp and online course.

I am a highly-motivated learner to keep me stay relevant, have good analytical thinking, have creative problem solving skill, and able to work in team. During college life, I developed my soft skill by having experienced in leading a commitee, participating in international competition (Model United Nation and business case competition), and also participating in extra-campus organization, such as AIESEC and StudentsCatalyst. ”

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

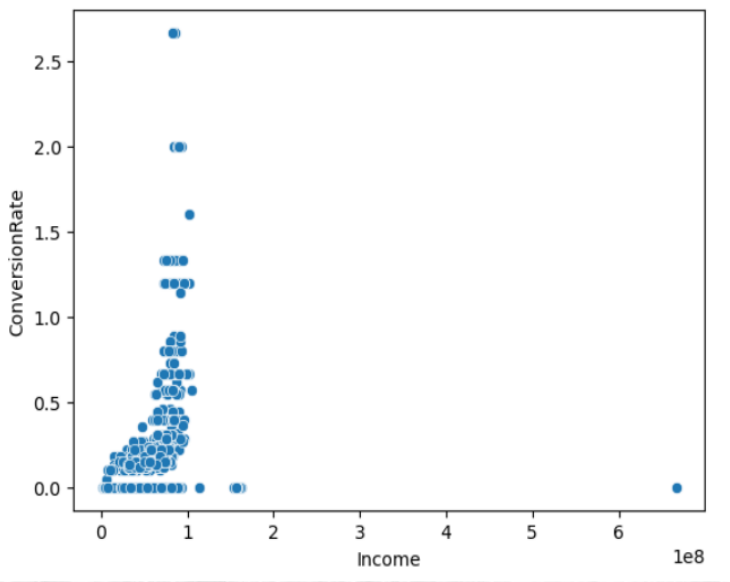
- Lakukan Feature Engineering dengan menghitung conversion rate dengan definisi (#response / #visit). Tidak hanya conversion rate, namun juga cari feature lain yang representatif, contohnya seperti umur, jumlah anak, total pengeluaran, total transaksi, dll.

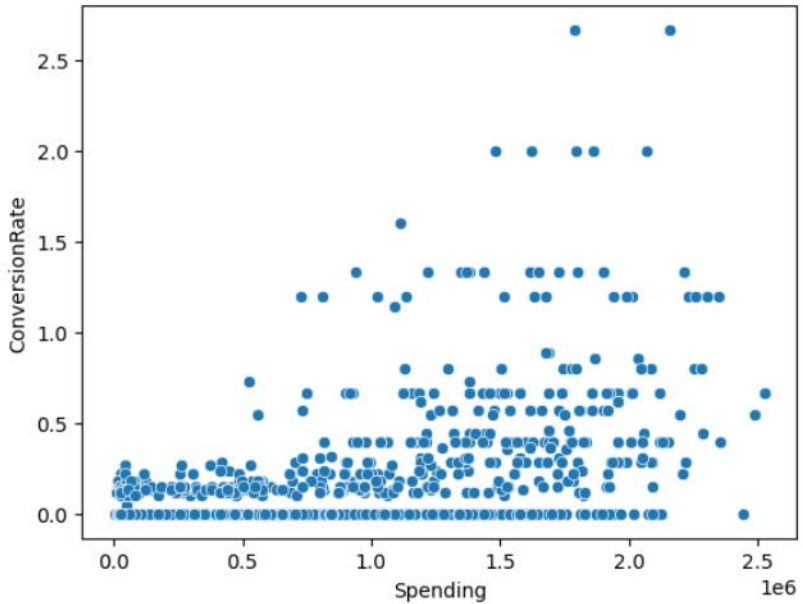
- Tulislah ***Exploration Data Analysis*** (EDA) yang sudah kamu lakukan, mulai dari plot yang kamu buat hingga analisis interpretasinya. Tuliskan pula insight yang dapat dijadikan rekomendasi (jika ada).

- **Source code** yang sudah kamu buat, dapat ditampilkan dan berikan link untuk mengakses file tersebut. Contohnya seperti di pojok kanan bawah.

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing

Conversion Rate Analysis Based on Income, Spending and Age

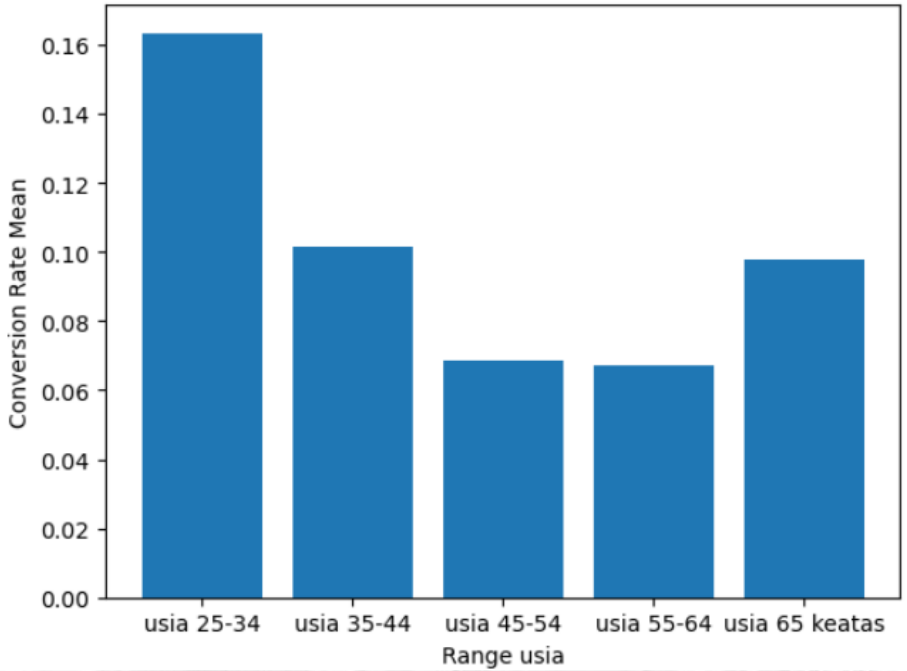




Semakin besar Income dan Spending, semakin besar kecenderungan conversion rate semakin tinggi

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing

Conversion Rate Analysis Based on Income, Spending and Age



Semakin besar usia, conversion rate cenderung semakin kecil.

Pengecualian terjadi pada usia 65 tahun keatas. Hal ini dikarenakan perhitungan conversion rate adalah response dibagi visit website sedangkan usia 65 tahun keatas yang melakukan transaksi cenderung jarang membuka website

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing

Data Cleaning & Preprocessing

- Pada tahap ***cleaning data***, tunjukan ***null*** atau ***missing value*** serta ***duplicated value*** pada dataset, serta cara penyelesaiannya.

- Selanjutnya untuk data preprocessing, tunjukan bahwa data sudah dilakukan proses ***feature encoding*** dan ***feature standardisation***.

- **Source code** yang sudah kamu buat, dapat ditampilkan dan berikan link untuk mengakses file tersebut. Contohnya seperti di pojok kanan bawah.

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing

TAHAP DATA CLEANING

```
Jumlah Missing Data:
Unnamed: 0              0
ID                      0
Year_Birth              0
Education               0
Marital_Status          0
Income                 24
Kidhome                 0
Teenhome                0
Dt_Customer             0
Recency                 0
MntCoke                 0
MntFruits               0
MntMeatProducts         0
MntFishProducts         0
MntSweetProducts        0
MntGoldProds            0
NumDealsPurchases       0
NumWebPurchases         0
NumCatalogPurchases     0
NumStorePurchases       0
NumWebVisitsMonth       0
AcceptedCmp3            0
AcceptedCmp4            0
AcceptedCmp5            0
AcceptedCmp1            0
AcceptedCmp2            0
Complain                0
Z_CostContact           0
Z_Revenue               0
Response                0
Age                     0
Children                0
Spending                0
Transaction             0
Total_Accepted          0
ConversionRate          0
Age_Segment             0
Year_Customer           0
dtype: int64
-------------------
Jumlah Duplicate Data:
0
```

Cleaning Rows:
Jumlah data yang hilang tidak signifikan, hanya 24 baris dari 2240 baris (hanya sekitar 1% saja), itu pun hanya terdapat pada 1 kolom saja. Maka dari itu, baris yang terdapat data hilang dihapus saja.

Cleaning Columns:
Kolom yang dihapus adalah:
- Kolom yang nilainya tidak berhubungan dengan keputusan pembelian, spt ID
- Kolom yang sudah dilakukan feature enguneer, seperti Age, Year_Birth, Dt_Customer, AcceptedCmp1, MntFruits, dsb

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing

TAHAP FEATURE ENCODING

Encode Label Segmen Umur

Original Value	Encode Value
usia 0-17	1
usia 18-24	2
usia 25-34	3
usia 35-44	4
usia 45-54	5
usia 55-64	6
usia 65 tahun keatas	7

Encode Label Education

Original Value	Encode Value
SMA	3
D3	6
S1	7
S2	9
S3	11

Encode Label Marital Status

Original Value	Encode Value
Menikah	1
Bertunangan	1
Lajang	0
Janda	0
Cerai	0
Duda	0

- Untuk kolom Education, angka pada encoding ditentukan dari jumlah tahun studi dari awal masuk SMA sampai pendidikan terakhir. Nilai encoding bukan 1-5 karena jarak waktu studi antara lulusan SMA – lulusan D3 dengan lulusan D3 - lulusan S1 tidaklah sama
- Untuk kolom Marital Status, angka 1 menunjukkan bahwa customer memiliki pasangan sedangkan angka 0 berarti tidak memiliki pasangan.

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing

TAHAP FEATURE STANDARDIZATION

	Education	Marital_Status	Income	Recency	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	Complain	Response	Children	Spending	Transaction	Total_Accepted	ConversionRate	Age_Segment	Year_Customer
0	7	0	58138000.0	58	3	8	10	4	7	0	1	0	1617000	25	0	0.000000	7	2012
1	7	0	46344000.0	38	2	1	1	2	5	0	0	2	27000	6	0	0.000000	7	2014
2	7	1	71613000.0	26	1	8	2	10	4	0	0	0	776000	21	0	0.000000	6	2013
3	7	1	26646000.0	26	2	2	0	4	6	0	0	1	53000	8	0	0.000000	4	2014
4	11	1	58293000.0	94	5	5	3	6	5	0	0	1	422000	19	0	0.000000	4	2014
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2235	7	1	61223000.0	46	2	9	3	4	5	0	0	1	1341000	18	0	0.000000	6	2013
2236	11	1	64014000.0	56	7	8	2	5	7	0	0	3	444000	22	1	0.133333	7	2014
2237	7	0	56981000.0	91	1	2	3	13	6	0	0	0	1241000	19	1	0.153846	4	2014
2238	9	1	69245000.0	8	2	6	5	10	3	0	0	1	843000	23	0	0.000000	7	2014
2239	11	1	52869000.0	40	3	3	1	4	7	0	1	2	172000	11	0	0.000000	7	2012

	Education	Marital_Status	Income	Recency	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	Complain	Children	Spending	Transaction	Total_Accepted	ConversionRate	Age_Segment	Year_Customer
0	-0.532874	-1.348829	0.234063	0.310532	0.351713	1.428553	2.504712	-0.554143	0.693232	-0.097812	-1.264803	1.675488	1.319446	-0.439265	-0.342481	1.383124	-1.500343
1	-0.532874	-1.348829	-0.234559	-0.380509	-0.168231	-1.125881	-0.571082	-1.169518	-0.131574	-0.097812	1.405806	-0.962358	-1.157987	-0.439265	-0.342481	1.383124	1.417393
2	-0.532874	0.741384	0.769478	-0.795134	-0.688176	1.428553	-0.229327	1.291982	-0.543978	-0.097812	-1.264803	0.280250	0.797881	-0.439265	-0.342481	0.504396	-0.041475
3	-0.532874	0.741384	-1.017239	-0.795134	-0.168231	-0.760962	-0.912837	-0.554143	0.280829	-0.097812	0.070501	-0.919224	-0.897205	-0.439265	-0.342481	-1.253059	1.417393
4	1.577669	0.741384	0.240221	1.554407	1.391603	0.333796	0.112428	0.061232	-0.131574	-0.097812	0.070501	-0.307044	0.537099	-0.439265	-0.342481	-1.253059	1.417393
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2235	-0.532874	0.741384	0.356642	-0.104093	-0.168231	1.793473	0.112428	-0.554143	-0.131574	-0.097812	0.070501	1.217598	0.406708	-0.439265	-0.342481	0.504396	-0.041475
2236	1.577669	0.741384	0.467539	0.241428	2.431492	1.428553	-0.229327	-0.246455	0.693232	-0.097812	2.741110	-0.270546	0.928273	1.033369	0.202108	1.383124	1.417393
2237	-0.532874	-1.348829	0.188091	1.450751	-0.688176	-0.760962	0.112428	2.215044	0.280829	-0.097812	-1.264803	1.051696	0.537099	1.033369	0.285891	-1.253059	1.417393
2238	0.522398	0.741384	0.675388	-1.417072	-0.168231	0.698715	0.795937	1.291982	-0.956381	-0.097812	0.070501	0.391404	1.058664	-0.439265	-0.342481	1.383124	1.417393
2239	1.577669	0.741384	0.024705	-0.311405	0.351713	-0.396043	-0.571082	-0.554143	0.693232	-0.097812	1.405806	-0.721800	-0.506031	-0.439265	-0.342481	1.383124	-1.500343

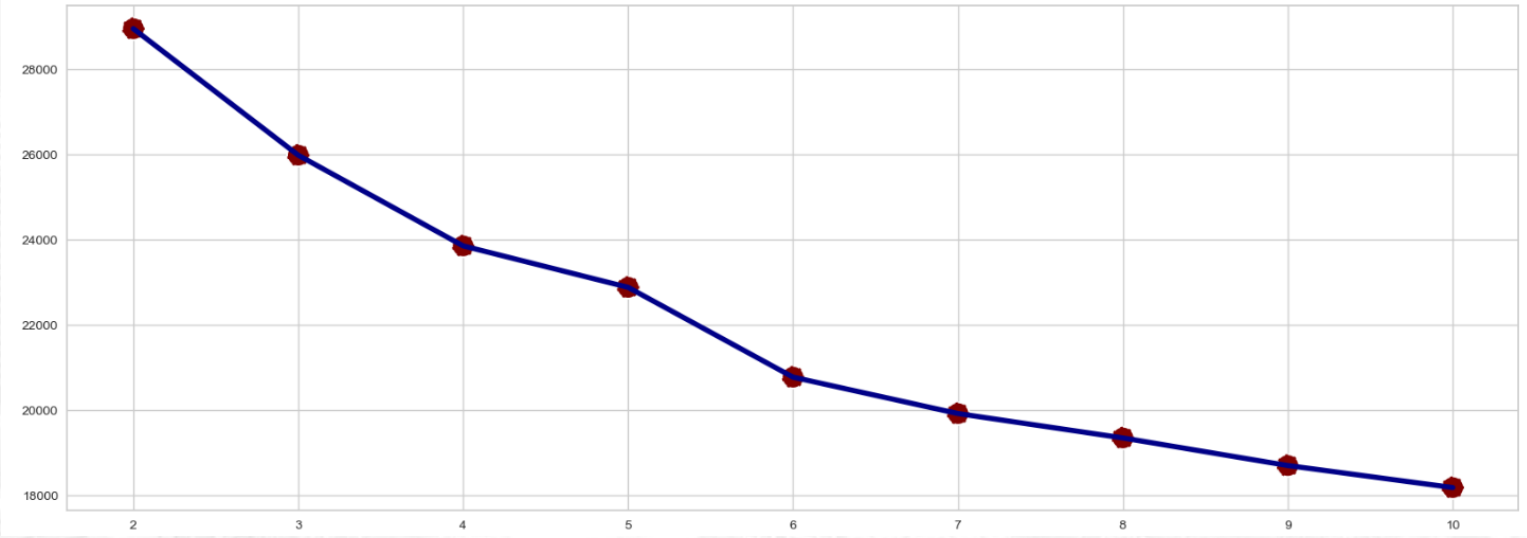
Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing

Data Modeling

- Tunjukan visualisasi ***Elbow Method*** menggunakan ***K-Means Clustering*** dan hasil evaluasinya menggunakan ***Silhouette Score***, serta buatkan lah hasil interpretasinya.

Untuk selengkapnya, dapat melihat jupyter notebook disini

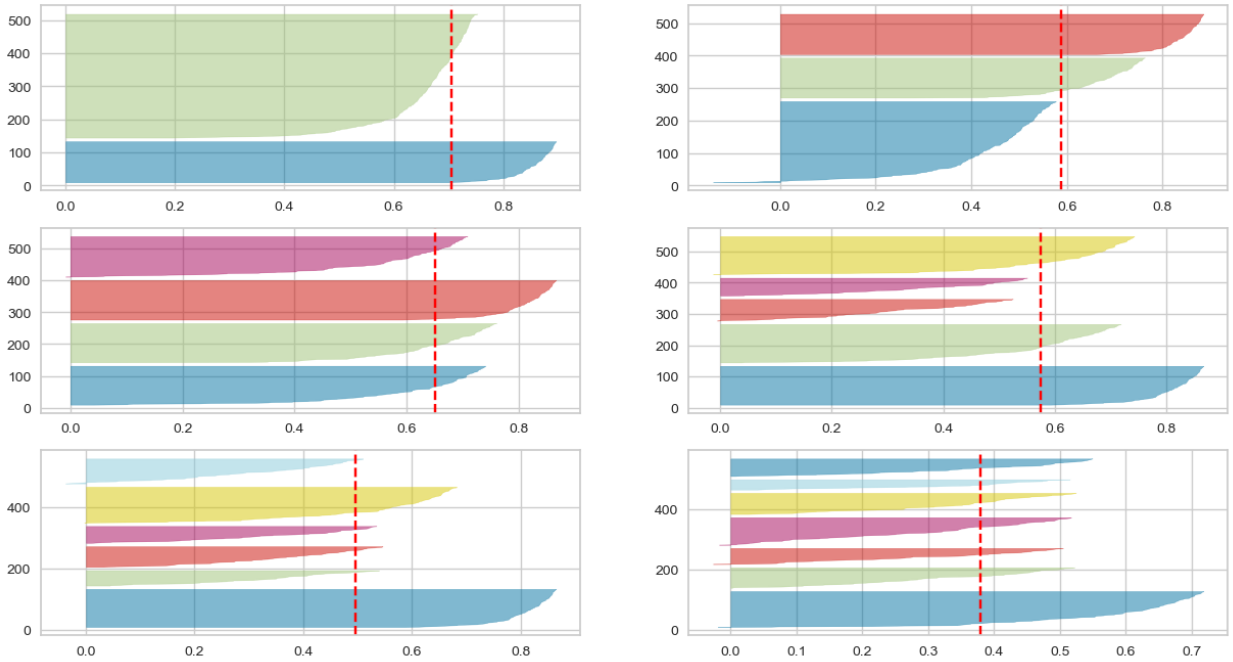
Visualisasi Elbow Method



Visualisasi diatas menunjukkan bahwa jumlah cluster yang ideal adalah 4 karena setelah poin ke-4, selisih perubahan nilai inertia tidak terlalu besar

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing

Analisa Shiluette Score



Visualisasi diatas menunjukkan bahwa jumlah cluster yang ideal adalah 4 karena pada cluster berjumlah 4:

1. Ujung akhir shiluette melebihi garis tiitik-titik merah
2. Nilai negative sangat minim serta
3. Ukuran shiluette tidak jauh berbeda satu sama lain

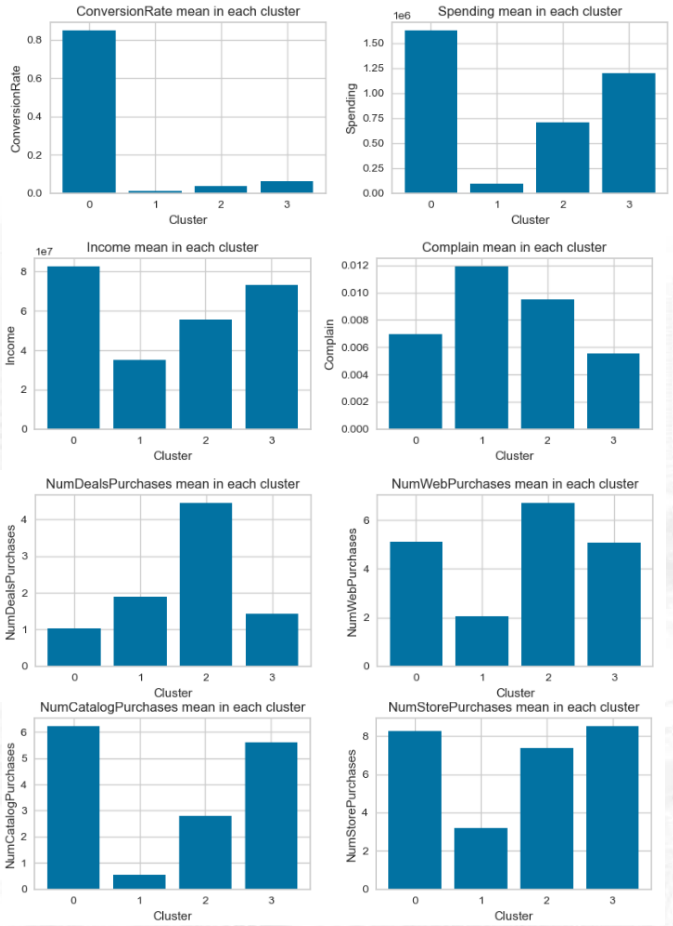
Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing

Customer Personality Analysis for Marketing Retargeting

- Tunjukan visualisasi analisis dari EDA dengan menggunakan ***hasil cluster*** yang sudah didapat. Buatlah rekomendasi bisnis yang dapat dilakukan dari analisis tersebut.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting



Visualisasi clustering dan Analisanya

Cluster 0 :
Sasaran Campaign

Memiliki keunggulan dalam conversionrate, spending, dan income, serta paling sedikit dalam menerima diskon sehingga layak menjadi sasaran campaign

Cluster 1 :
Kumpulan complainer

Memiliki rata-rata complain tertinggi namun conversionrate dan spending sangat rendah Sangat tidak direkomendasikan untuk ditargetkan campaign karena hanya akan mempersulit CS membalas complaint dan menambah cost campaign

Cluster 2 :
Pencari Promo Online

Memiliki keunggulan dalam NumDealsPurchase dan NumWebPurchase, namun sedikit yang berbelanja di store

Cluster 3 :
Target Cadangan

Memiliki keunggulan dalam NumStorePurchase dan jarang menerima diskon. Berada pada urutan kedua dalam hal ConversionRate, Spending, dan Income. Relatif layak dijadikan target campaign meski tidak sebagus cluster 0

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing

Customer Personality Analysis for Marketing Retargeting

Pemilihan Cluster untuk Re-targetting

	cluster	Age_Segment	Income	Recency	Spending	Complain	ConversionRate	Total_Accepted	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases
0	0	5.333333	8.245504e+07	48.034722	1.621257e+06	0.006944	0.847735	2.256944	1.041667	5.118056	6.208333	8.243056
1	1	5.192460	3.500966e+07	48.999008	9.423611e+04	0.011905	0.011991	0.084325	1.886905	2.048611	0.536706	3.197421
2	2	5.786667	5.544722e+07	49.243810	7.081924e+05	0.009524	0.034738	0.249524	4.441905	6.693333	2.800000	7.363810
3	3	5.536178	7.329658e+07	49.074212	1.196711e+06	0.005566	0.061995	0.222635	1.419295	5.077922	5.591837	8.495362

Cluster yang dipilih untuk retargeting adalah cluster 0 karena cluster tersebut memiliki keunggulan dalam nilai rata-rata Spending, Income, Conversion Rate, dan Total_Accepted.

	cluster	Jumlah penerima campaign	Jumlah Anggota Cluster
0	0	83	144
1	1	86	1008
2	2	76	525
3	3	88	539

Hal ini sesuai dengan analisa Response Rate dimana cluster 0 memiliki response rate paling tinggi

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing

Perhitungan Business Impact dari Klusterisasi

Jika diasumsikan biaya (cost) untuk melakukan campaign adalah 3 dollar, dan pendapatan (revenue) jika user menerima campaign adalah 11 dollar, maka perhitungan potensi penambahan profit jika marketing campaign hanya menyasar segmen 0 adalah sebagai berikut:

Metrics	Blast Marketing	Retargeting ke cluster 0
Jumlah target	2216	144
Cost Campaign	$6648	432
Jumlah yang meresponse campaign	333	83
Revenue dari Campaign	$3663	913
Profit Campaign	- $2985	$481

Clusterisasi berhasil memperbaiki profitabilitas perusahaan dari yang awalnya **merugi 2985 dollar** menjadi **profit 481 dollar**

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/1J-2NWwrTrKxE_jyqf_u4H5Iusn8x7i_e?usp=sharing