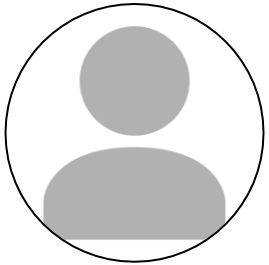
Predict Customer Personality to boost marketing campaign by using Machine Learning

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



Created by:
Ferry Irwanto
ferryirwanto89@gmail.com
linkedin.com/in/ferryirwanto

"Data Analyst with hands-on experience in financial risk analysis, KPI tracking, and dashboard development. Proficient in SQL, Python, and BI tools to transform raw data into actionable insights. Skilled at managing multiple projects independently, delivering accurate business solutions under tight deadlines and improving operational efficiency."

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).

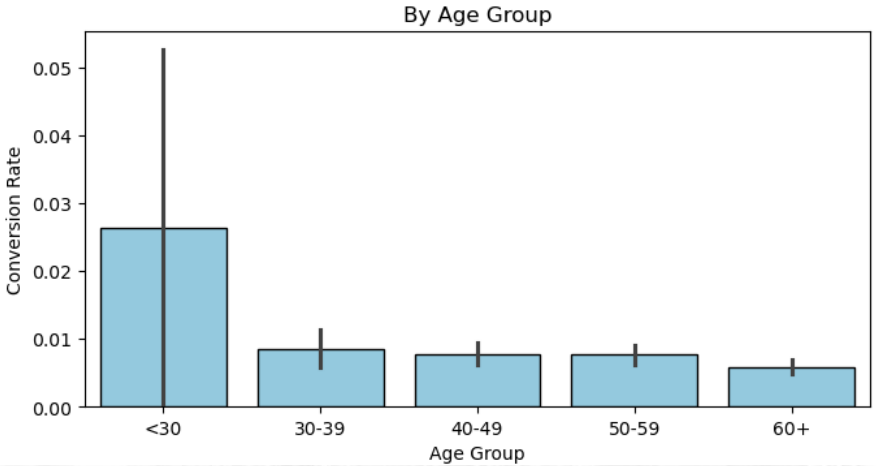
- **Source code** yang sudah kamu buat, dapat ditampilkan dan berikan link untuk mengakses file tersebut. Contohnya seperti di pojok kanan bawah.

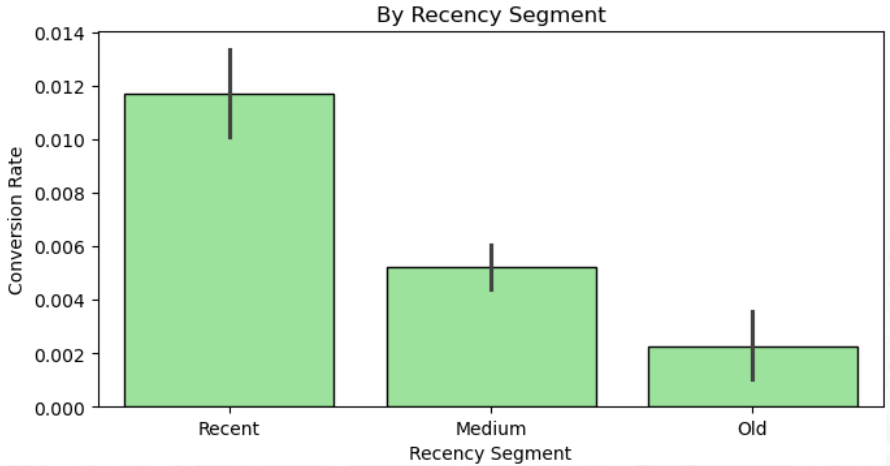
Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Engineered

Feature Names	Description
Customer Value Index	Mengukur seberapa besar nilai belanja customer
Family Size	Bisa menentukan apakah campaign cocok untuk kebutuhan rumah tangga besar/kecil
Engagement Score	Gabungan dari aktivitas belanja & kunjungan
Recency Segment	Berapa lama sejak pembelian terakhir
Income per Capita	Bisa memengaruhi seberapa responsive mereka terhadap campaign diskon/premium
Loyalty Indicator	Indikator banyakanya ikut campaign sebelumnya

Conversion rate by user segment

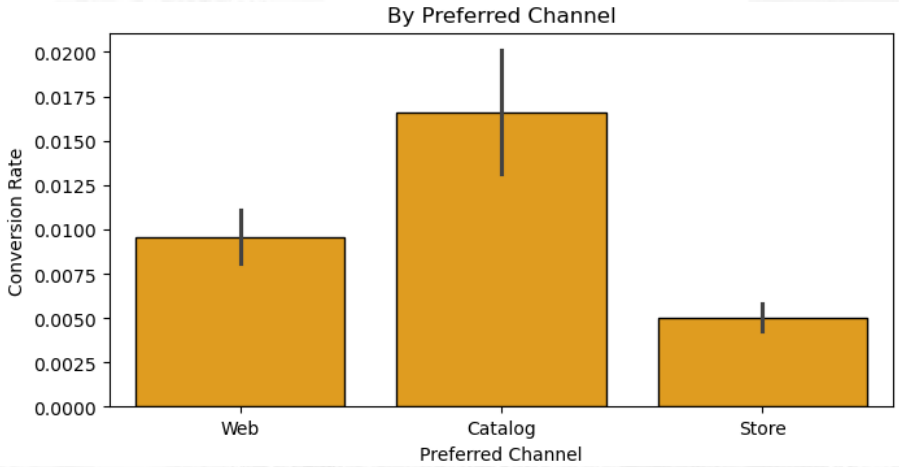


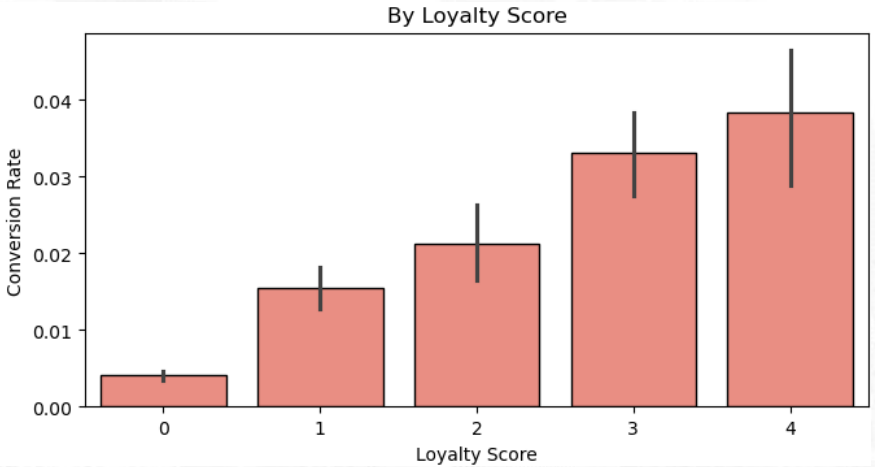


- Customer usia <30 menunjukkan conversion rate lebih tinggi dibanding semua kelompok lain.
- Setelah usia 30 tahun, conversion rate turun drastis dan relatif stabil rendah.
- Kesimpulan: ada hubungan signifikan antara umur dan conversion rate dimana semakin muda customer, semakin besar kemungkinan mereka merespons campaign.

- Customer yang Recent (belanja dalam 30 hari terakhir) jauh lebih responsif daripada yang Medium/Old.
- Ini konsisten dengan teori marketing: customer aktif lebih mudah diaktivasi kembali.

Conversion rate by user segment

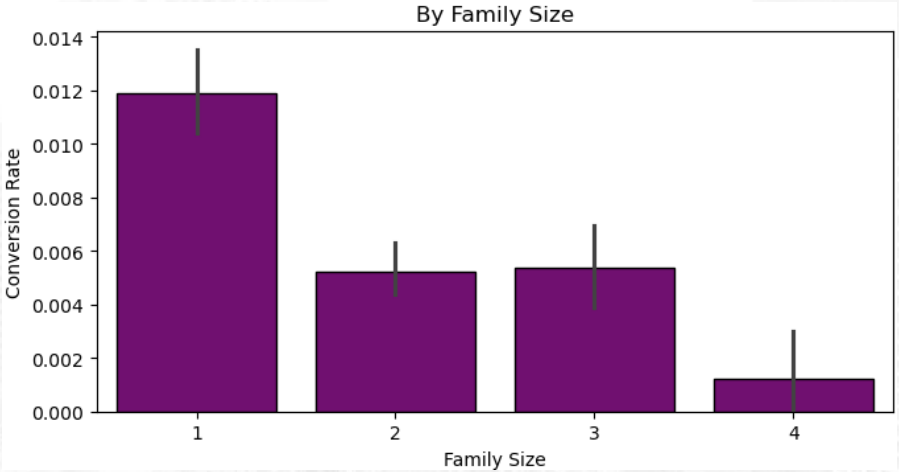




- Catalog buyers punya conversion rate paling tinggi, disusul Web, lalu Store paling rendah.
- Artinya channel komunikasi & promosi perlu diarahkan lebih kuat ke Catalog dan Web users.

- Ada tren positif: semakin sering customer menerima/ikut campaign sebelumnya, semakin besar kemungkinan mereka merespons campaign selanjutnya.
- Customer yang sudah pernah ikut 3–4 campaign sebelumnya adalah target terbaik.

Conversion rate by user segment



- Conversion rate tertinggi ada pada single customer (Family Size = 1).
- Semakin besar family size, cenderung lebih rendah respons campaign.
- Ini bisa jadi karena rumah tangga besar lebih price-sensitive atau prioritas belanja berbeda.

Kesimpulan Utama

- Umur <30 adalah faktor signifikan: customer muda jauh lebih responsif.
- Selain umur, faktor penting lain: recent activity, catalog/web buyers, high loyalty, single household → mereka adalah jenis user paling potensial untuk ditarget campaign.

Rekomendasi Strategi Utama

- Segmentasi Target: Utamakan usia <30, recent buyers, catalog/web buyers, dan high loyalty customers.
- Channel Strategy: Dorong campaign melalui digital & catalog channel, alokasikan lebih sedikit ke store-only buyers.
- Personalization: Bedakan campaign untuk single vs family customer agar lebih relevan.
- Retention Focus: Berikan campaign tepat setelah pembelian untuk meningkatkan repeat purchase.

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

Data Cleaning

```
df = df.drop_duplicates()
print("Missing values before handling:\n", df.isnull().sum().sort_values(ascending=False))
```

✓ 0.0s

```
Missing values before handling:
 Income                24
Income_per_Capita      24
ID                      0
Conversion_Rate         0
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntCoke	...	Conversion_Rate	Total_Spending	Family_Size	Recency_Segment	Income_per_Capita
10	1994	1983	S1	Menikah	NaN	1	0	2013-11-15	11	5000	...	0.000000	19000	2	Recent	NaN
27	5255	1986	S1	Lajang	NaN	1	0	2013-02-20	19	5000	...	0.000000	637000	2	Recent	NaN
43	7281	1959	S3	Lajang	NaN	0	0	2013-11-05	80	81000	...	0.000000	186000	1	Medium	NaN
48	7244	1951	S1	Lajang	NaN	2	1	2014-01-01	96	48000	...	0.000000	124000	4	Old	NaN
58	8557	1982	S1	Lajang	NaN	1	0	2013-06-17	57	11000	...	0.000000	46000	2	Medium	NaN
71	10629	1973	D3	Menikah	NaN	1	0	2012-09-14	25	25000	...	0.000000	109000	2	Recent	NaN
90	8996	1957	S3	Menikah	NaN	2	1	2012-11-19	4	230000	...	0.000000	603000	4	Recent	NaN

- Data duplikat akan langsung dihapus karena tidak ada relevansi terhadap model dengan menggunakan .drop_duplicates()
- Terdapat 24 baris missing value pada fitur Income dan Income_per_capita
- Untuk mengatasinya dilakukan imputasi, Solusi ini digunakan karena fitur dianggap penting secara logika bisnis
- Alasan lain Adalah ketika di crosscheck dengan Year_Birth atau dari fitur lainnya (Age), hasil yang dilihat menandakan bahwa memang ada masalah pada input data

- Imputasi dilakukan dengan mengisi median pada Age_Group yang sama
- Khusus Income_per_Capita didasarkan dengan [Income/family_size] jika family_size tidak 0

Data Preprocessing

MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	...	Marital_Status_Lajang	Marital_Status_Menikah	Recency_Segment_Old	Recency_Segment_Recent
546000	172000	88000	88000	...	True	False	False	False
6000	2000	1000	6000	...	True	False	False	False
127000	111000	21000	42000	...	False	False	False	True
20000	10000	3000	5000	...	False	False	False	True
118000	46000	27000	15000	...	False	True	True	False

Before Encoding and Standardize

After Encoding and Standardize

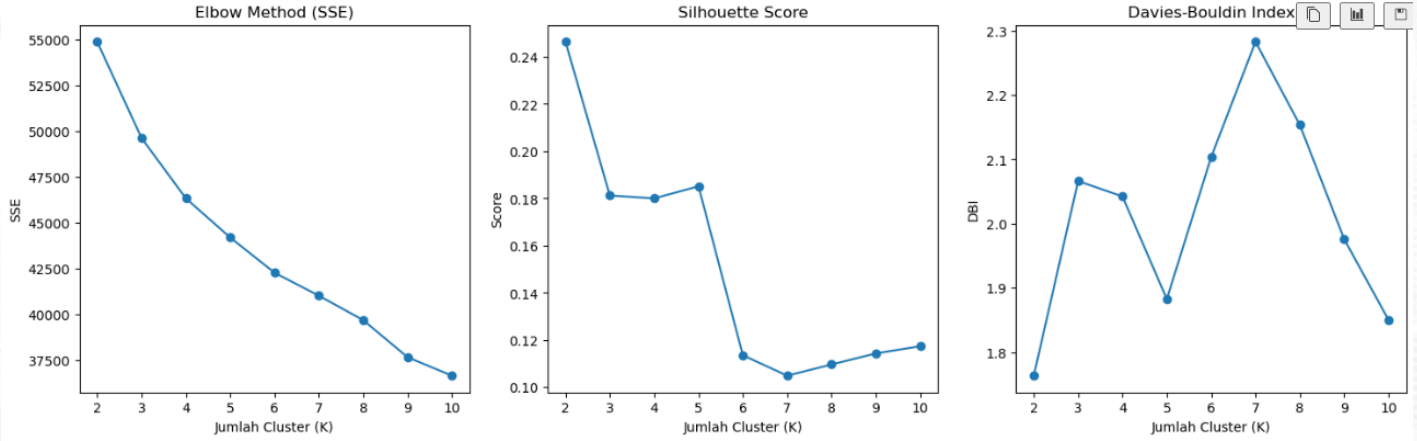
MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	...	Marital_Status_Lajang	Marital_Status_Menikah	Recency_Segment_Old	Recency_Segment_Recent
1.679702	2.462147	1.476500	0.843207	...	True	False	False	False
-0.713225	-0.650449	-0.631503	-0.729006	...	True	False	False	False
-0.177032	1.345274	-0.146905	-0.038766	...	False	False	False	True
-0.651187	-0.503974	-0.583043	-0.748179	...	False	False	False	True
-0.216914	0.155164	-0.001525	-0.556446	...	False	True	True	False

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

Elbow Method & Silhouette Score




```
score = silhouette_score(df_final.drop("Cluster", axis=1), df_final["Cluster"])
print(f"Silhouette Score untuk k={k}: {score:.4f}")
```

✓ 0.0s

```
Silhouette Score untuk k=5: 0.1851
```

K=5 adalah pilihan terbaik karena:

- Didukung Elbow Method (SSE melandai setelah 5).
- DBI relatif rendah (cluster cukup terpisah).
- Silhouette Score memang tidak tinggi (<0.25), tapi ini wajar di data customer behavior yang kompleks.

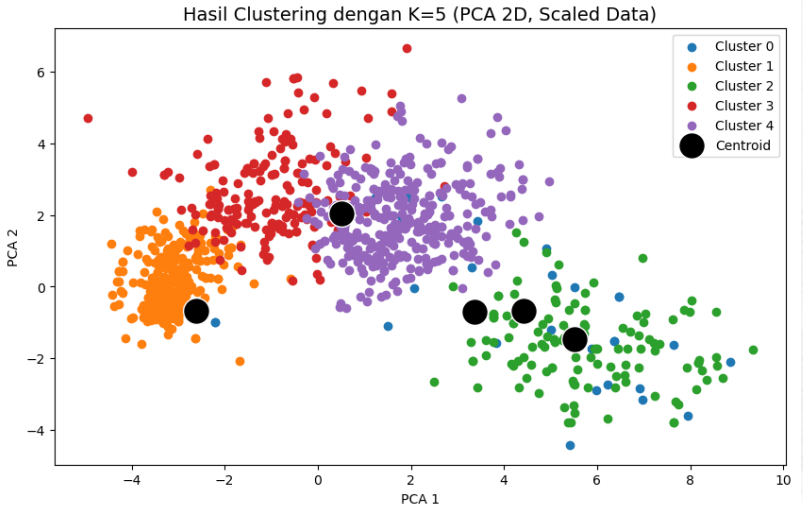
- Untuk nilai Silhouette score masih kurang optimal (<=0.25)

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

Interpretasi Cluster



- Cluster terlihat terpisah sebagian
  - Warna orange (Cluster 1) cukup jelas terpisah dari cluster lain.
  - Warna hijau (Cluster 2) juga punya area dominan sendiri.
  - Warna merah (Cluster 3) dan ungu (Cluster 4) agak overlap → artinya pola konsumsi mereka mirip.
  - Warna biru (Cluster 0) tersebar tipis dan sebagian bercampur dengan hijau → bisa jadi ini cluster kecil atau "outlier group".
- Posisi centroid masih logis
  - Centroid (titik hitam) ada di tengah-tengah distribusi tiap cluster.
  - Tidak ada centroid yang "nganggur" jauh dari titik-titik → ini indikasi clustering bekerja sesuai data.
- Normal untuk data real world
  - Overlap antara merah & ungu itu wajar menandakan bahwa data aslinya memang ada customer yang perilakunya mirip di dua segmen.

Seleksi cluster untuk retargetting

Cluster	Dominan Usia	Recency (Perilaku)	Income	Spending	Channel Utama	Interpretasi Utama
0	60+	Rendah (baru belanja)	Sedikit di atas rata-rata	Sedikit di atas rata-rata	Store	Senior baru aktif belanja, cocok untuk program loyalitas
1	50–59	Tinggi (lama tidak belanja)	Di bawah rata-rata	Di bawah rata-rata	Store	Hampir pensiun, dorman → butuh reaktivasi dengan promo
2	60+	Rendah (baru belanja)	Di atas rata-rata	Jauh di atas rata-rata	Store	Senior mapan, daya beli tinggi → target produk premium
3	60+	Rendah (baru belanja)	Sekitar rata-rata	Sekitar rata-rata	Store	Senior aktif belanja, cocok untuk campaign massal
4	60+	Tinggi (lama tidak belanja)	Di atas rata-rata	Di atas rata-rata	Store	Senior mapan tapi dorman → campaign reaktivasi eksklusif

Kandidat terbaik biasanya:
- Income tinggi (≥ 1)
- Spending besar (≥ 1 atau 2)
- Recency tinggi (lama tidak belanja)

Dari tabel: Cluster 4 paling cocok → karena income + spending relatif tinggi, tapi sudah lama tidak belanja → cocok untuk retargeting.

Kalkulasi potential impact

```
Ringkasan Cluster
         Jumlah_Customer  Rata2_Spending
Cluster
0                     30            1.17
1                   1032           -0.85
2                    164            1.71
3                    612            0.23
4                    402            1.04

Target Cluster: 4
Potential Impact (10% konversi): 41.81
```

Cluster	Penjelasan
Cluster 0	Sangat kecil (30 orang), spending lumayan (1.17) artinya cluster ini kurang signifikan karena populasinya kecil.
Cluster 1	Jumlah customer sangat besar (1032), tapi spending rata-rata negatif/di bawah rata-rata artinya kelompok ini cenderung tidak aktif atau belanja sedikit.
Cluster 2	Spending rata-rata paling tinggi (1.71), tapi jumlah customer kecil (164). Ini adalah kelompok high spender tetapi niche.
Cluster 3	Jumlah cukup besar (612), spending rata-rata rendah (0.23).
Cluster 4	Jumlah customer menengah (402), spending rata-rata cukup tinggi (1.04).

Rekomendasi

Hasil Clustering	Rekomendasi
Cluster 0 (±30 customer)	Spending lumayan tapi kecil → tidak signifikan
Cluster 1 (±1000 customer)	Spending rendah → bukan prioritas
Cluster 2 (±160 customer)	Spending tinggi → loyalty/VIP program
Cluster 3 (±600 customer)	Spending sedang → promosi diskon/bundling
Cluster 4 (±400 customer)	Spending tinggi tapi lama tidak belanja → **target utama retargeting**