Predict Customer Personality to boost marketing campaign by using Machine Learning

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

Created by:
Feri Dwi Saputro
feridwisa95@gmai.com
linkedin.com/in/ferids55

"Saya seorang **mantan Electrical O&M** yang telah bekerja selama 3 tahun di PLTU dan kini ingin mengubah jalur karir ke bidang data karena lebih sesuai dengan passion saya yaitu analisis dan perhitungan numerik. Tools yang biasa saya gunakan adalah **SQL dan Python** untuk analisis data dan pemodelan. Saya tertarik pada penerapan data science dalam **bidang retail, marketing, atau mungkin esport**."

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 ”

Data Overview

Load Dataset

```
# read csv file into dataframe with first column as index
df_campaigns = pd.read_csv('data/marketing_campaign_data.csv', index_col=0)
# print first five rows
df_campaigns.head()
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntCoke	MntFruits	MntMeatProducts
0	5524	1957	S1	Lajang	58138000.0	0	0	04-09-2012	58	635000	88000	546000
1	2174	1954	S1	Lajang	46344000.0	1	1	08-03-2014	38	11000	1000	6000
2	4141	1965	S1	Bertunangan	71613000.0	0	0	21-08-2013	26	426000	49000	127000
3	6182	1984	S1	Bertunangan	26646000.0	1	0	10-02-2014	26	11000	4000	20000
4	5324	1981	S3	Menikah	58293000.0	1	0	19-01-2014	94	173000	43000	118000

- Terdiri dari 2240 baris dan 29 kolom
- Nama kolom dan tipe data sesuai
- Terdapat kolom dengan jumlah baris < total (Income)
- Perlu Feature Engineering karena beberapa kolom sejenis datanya

Information Data

```
# view attributus and datatypes
df_campaigns.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2240 entries, 0 to 2239
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
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

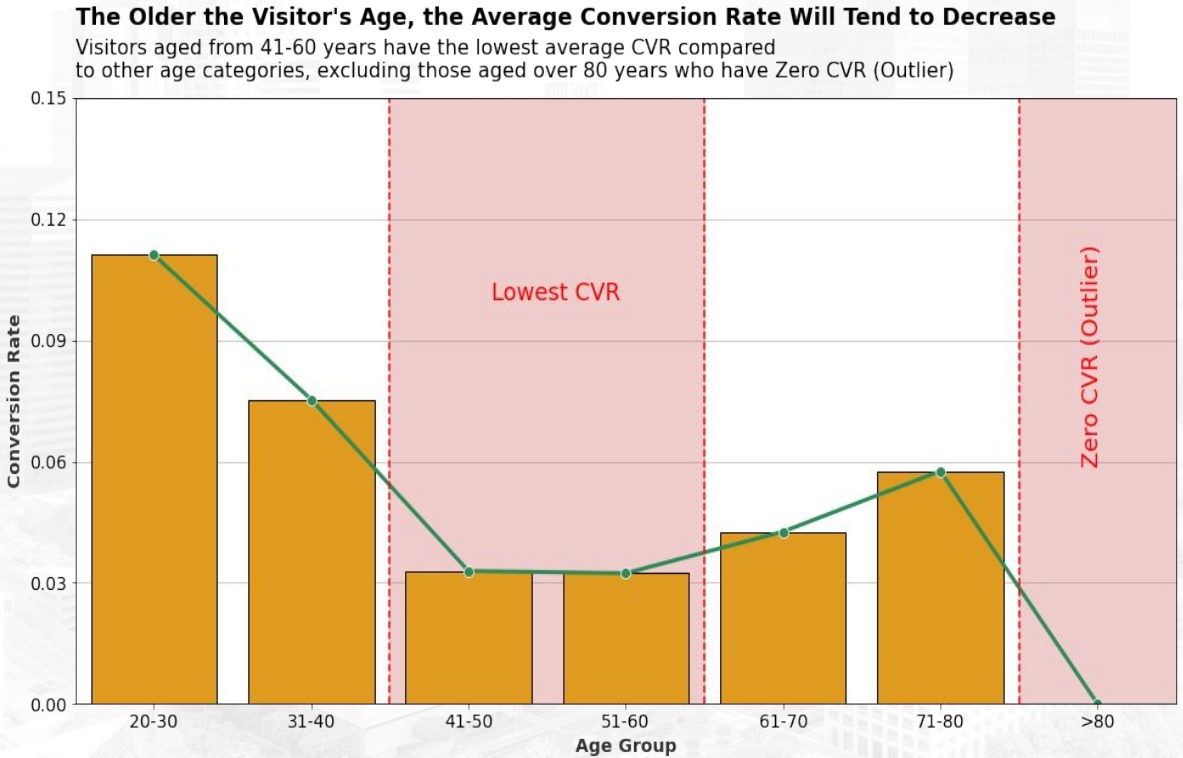
Feature Engineering

1. **Age** → (2022 - Year_Birth)
2. **Children** → (Kidhome + Teenhome)
3. **TotalSpent** → (MntCoke + MntFruits + MntMeatProducts + MntFishProducts + MntSweetProducts + MntGoldProds)
4. **NumOfTransactions** → (NumDealsPurchases + NumWebPurchases + NumCatalogPurchases + NumStorePurchases)
5. **NumOfAcceptedCmp** → (AcceptedCmp1 + AcceptedCmp2 + AcceptedCmp3 + AcceptedCmp4 + AcceptedCmp5)
6. **CVR** → (Response / NumWebVisitsMonth)

```
df_campaigns['Age'] = 2022 - df_campaigns['Year_Birth']
df_campaigns['Children'] = df_campaigns['Kidhome'] + df_campaigns['Teenhome']
df_campaigns['TotalSpent'] = df_campaigns.filter(regex='Mnt', axis=1).sum(axis=1)
df_campaigns['NumOfTransactions'] = df_campaigns.filter(regex='Purchase', axis=1).sum(axis=1)
df_campaigns['NumOfAcceptedCmp'] = df_campaigns.filter(regex='Cmp', axis=1).sum(axis=1)
df_campaigns['CVR'] = round(df_campaigns['Response']/df_campaigns['NumWebVisitsMonth'], 4)
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Age

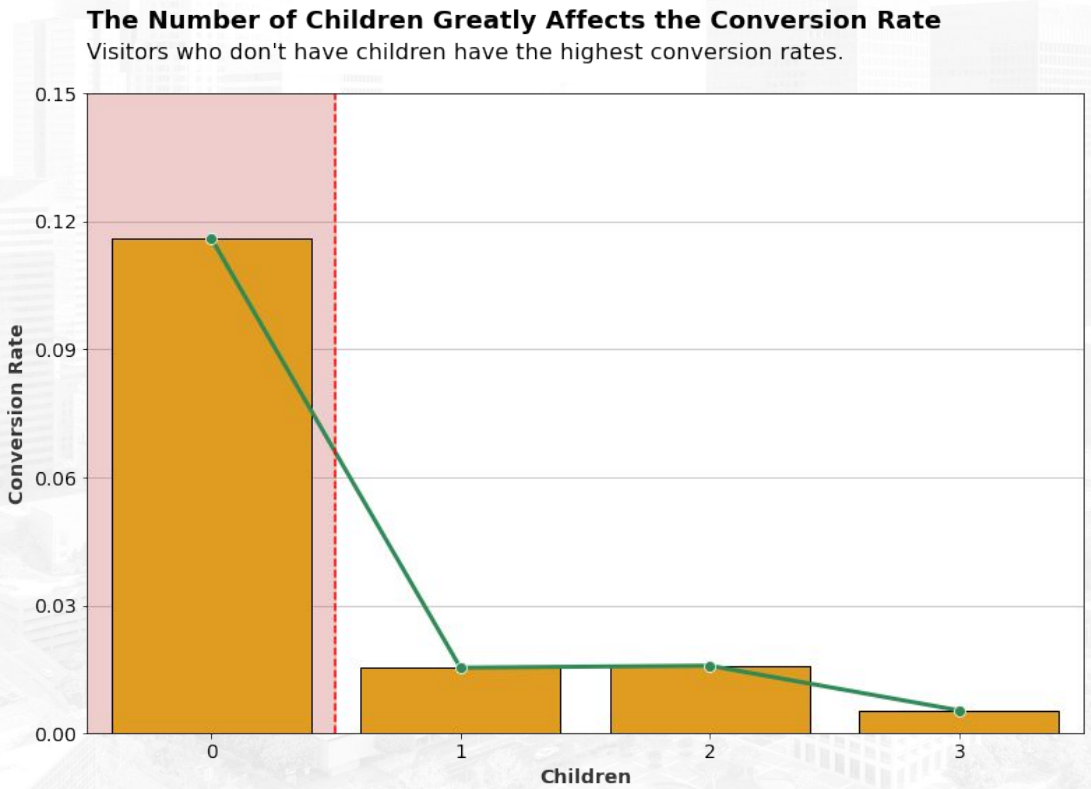


Insight:

Semakin tua umur pengunjung maka tingkat konversinya semakin menurun. Terlihat bahwa **pengunjung yang berumur 41 tahun hingga 60 tahun memiliki CVR paling rendah**. Selain itu, ditemukan adanya **outlier yaitu pengunjung yang berumur diatas 40 tahun** karena tidak memiliki CVR.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Children

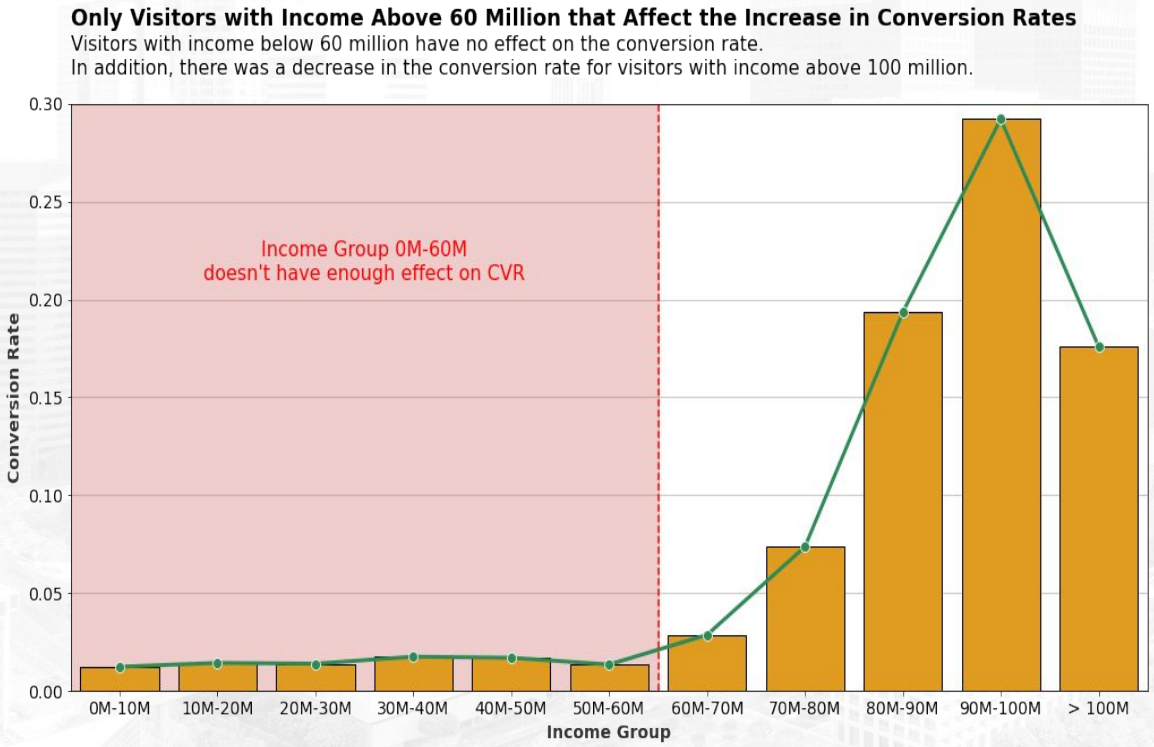


Insight:

Jumlah anak sangat berpengaruh terhadap CVR karena memiliki korelasi negatif. **Pengunjung yang belum memiliki anak memiliki CVR yang sangat tinggi.**

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Income

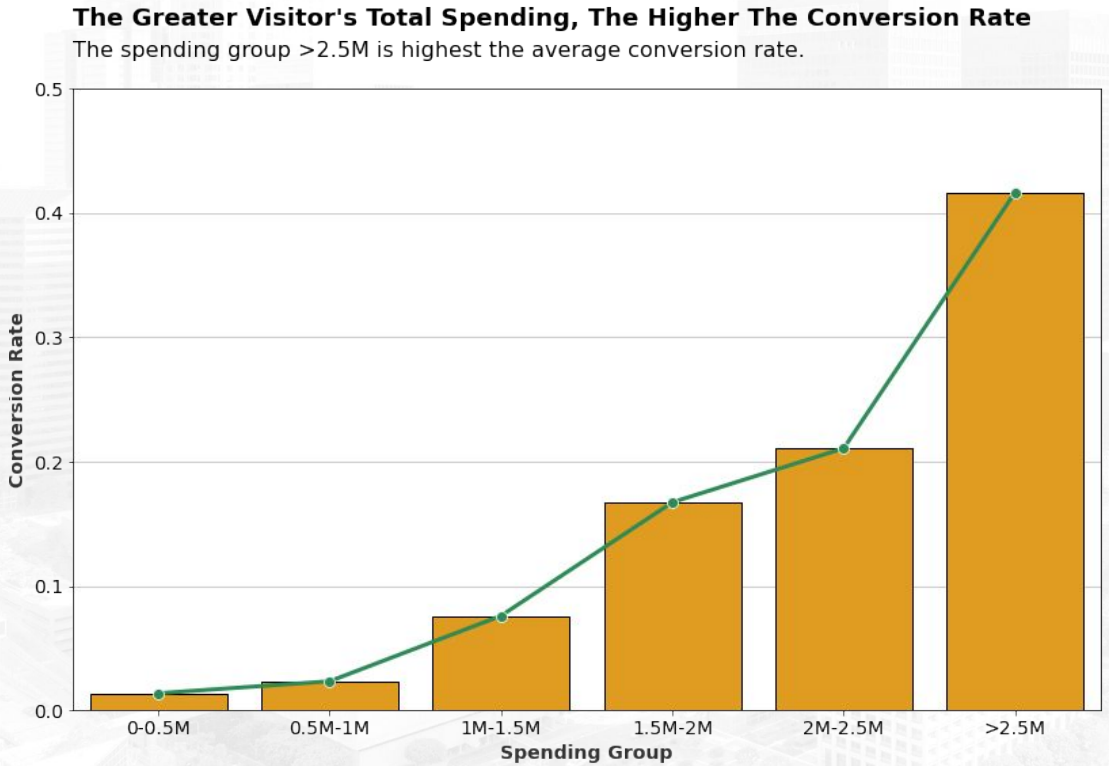


Insight:

Pengunjung dengan pendapatan dibawah 60 juta tidak berpengaruh terhadap CVR, sebaliknya pengunjung dengan pendapatan diatas 60 juta justru mempengaruhi CVR. **Pengunjung dengan pendapatan antara 91 juta sampai 100 juta memiliki CVR tertinggi** dan diatas kategori tersebut CVR mulai turun.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Total Spending

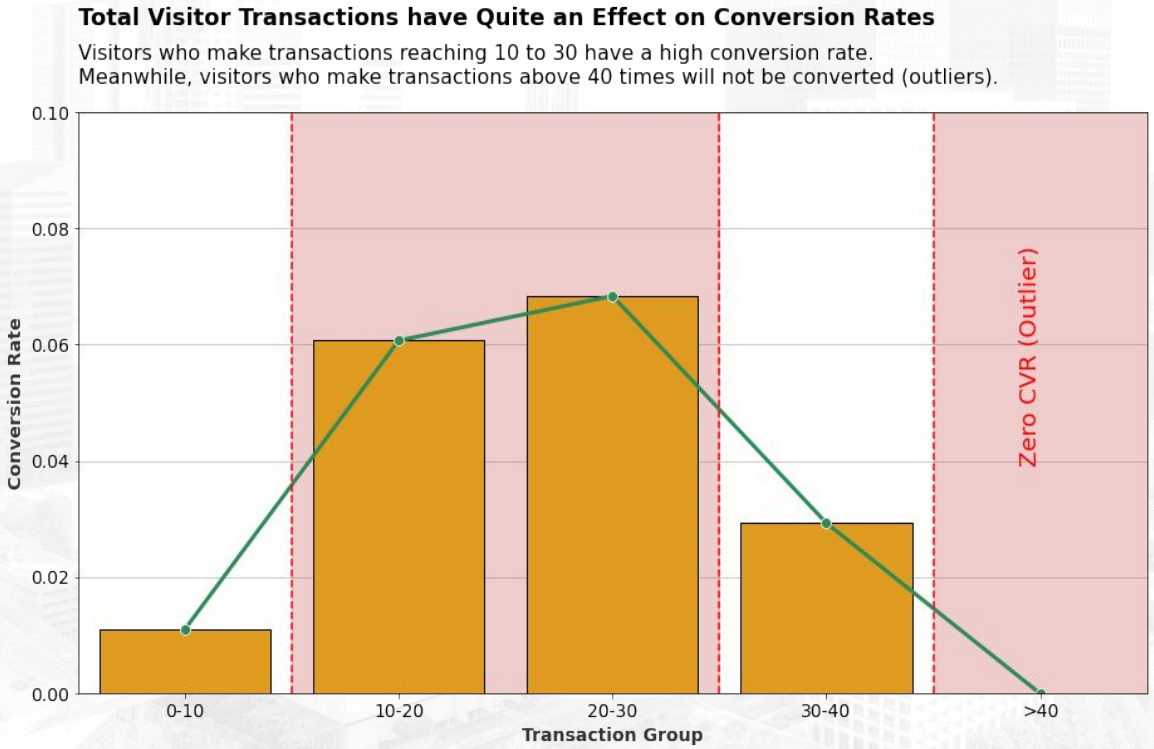


Insight:

Total pengeluaran memiliki korelasi positif kuat terhadap CVR. Semakin besar total pengeluaran maka semakin tinggi CVR nya.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Total Transaction

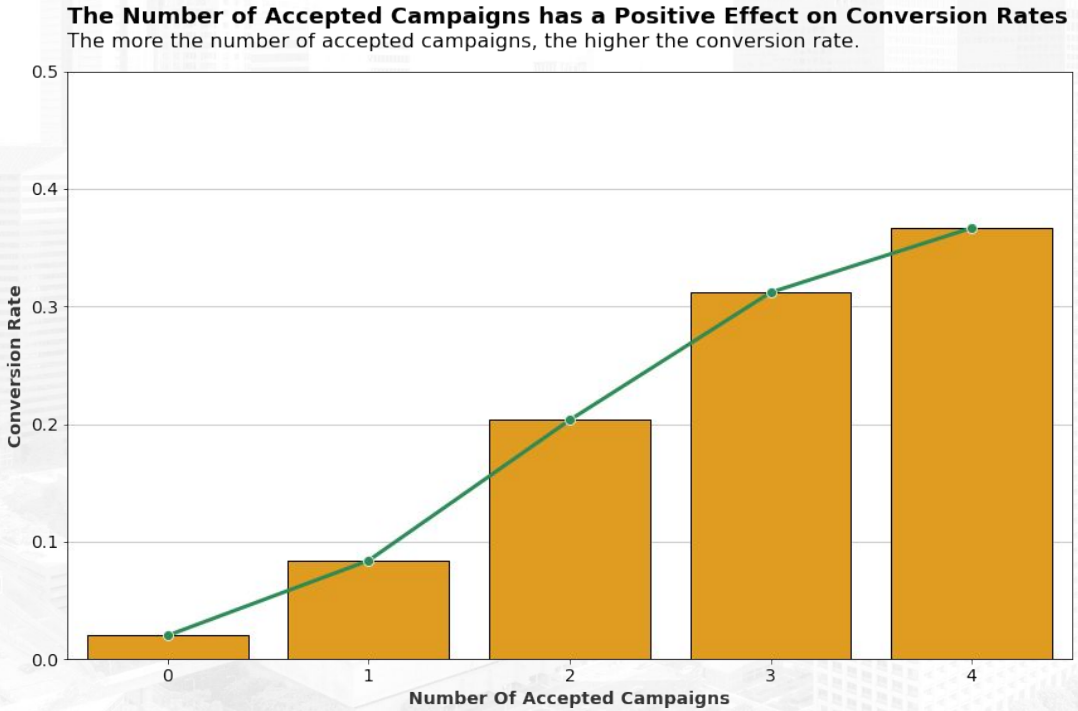


Insight:

Total Transaksi Pengunjung sedikit berpengaruh terhadap CVR. **Pengunjung yang telah melakukan transaksi mencapai 10 hingga 30 memiliki CVR yang paling tinggi**. Namun saat transaksi yang dilakukan diatas 40 kali sudah tidak akan terkonversi lagi (dianggap outlier).

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conversion Rate Analysis Based on Accepted Campaigns



Insight:

Jumlah campaign yang disetujui sangat berpengaruh pada CVR pengunjung yang ditunjukkan dengan korelasi posifif. **Semakin banyak jumlah campaign yang disetujui maka semakin tinggi CVR nya.**

Untuk selengkapnya, dapat melihat jupyter notebook disini

Missing Values

```
# view number of missing values
null_cols = df_campaign_prep.columns[df_campaign_prep.isnull().any()]
df_null = df_campaign_prep[null_cols].isnull().sum().to_frame().reset_index()
df_null.columns = ['kolom', 'jumlah']
df_null['persentase'] = round(df_null['jumlah']/len(df_campaign_prep) * 100, 3)
df_null
```

	kolom	jumlah	persentase
0	Income	24	1.071
1	IncomeGroup	24	1.071

Using Imputation Method

- Income → Median
- IncomeGroup → “50M-60M”

```
df_campaign_prep['Income'].fillna(df_campaign_prep['Income'].median(), inplace=True)
df_campaign_prep['IncomeGroup'].fillna('50M-60M', inplace=True)
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning

Duplicate Data → Drop

```
# remove duplicate rows
print(f'Jumlah data sebelum difilter duplikasi adalah {len(df_campaign_prep)} baris')
df_campaign_prep.drop('ID', axis=1, inplace=True)
df_campaign_prep.drop_duplicates(inplace=True)
print(f'Jumlah data setelah difilter duplikasi adalah {len(df_campaign_prep)} baris')
```

```
Jumlah data sebelum difilter duplikasi adalah 2240 baris
Jumlah data setelah difilter duplikasi adalah 2057 baris
```

Outlier Data → Filter

```
# view total rows before filtered
print(f'Jumlah baris sebelum difilter outlier adalah {len(df_campaign_prep)}')

# handle outlier using filtering based on EDA
filtered_entries = (df_campaign_prep['Income']>600000000) | (df_campaign_prep['Age']>80) | (df_campaign_prep['NumOfTransactions
df_campaign_prep = df_campaign_prep[~filtered_entries]

# view total rows after filtered
print(f'Jumlah baris setelah difilter outlier adalah {len(df_campaign_prep)}')
```

```
Jumlah baris sebelum difilter outlier adalah 2057
Jumlah baris setelah difilter outlier adalah 2049
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Preprocessing

Feature Encoding

1. Label Encoding → Education, AgeGroup, IncomeGroup, SpendingGroup, TransactionGroup

2. One Hot Encoding → Marital_Status

Before Encoding

	Education	Marital_Status	AgeGroup	IncomeGroup	SpendingGroup	TransactionGroup
0	S1	Lajang	61-70	50M-60M	1.5M - 2M	20-30
1	S1	Lajang	61-70	40M-50M	0 - 0.5M	0-10
2	S1	Bertunangan	51-60	70M-80M	0.5M - 1M	20-30
3	S1	Bertunangan	31-40	20M-30M	0 - 0.5M	0-10
4	S3	Menikah	41-50	50M-60M	0 - 0.5M	10-20

After Encoding

	Education	AgeGroup	IncomeGroup	SpendingGroup	TransactionGroup	MaritalStatus_Bertunangan	MaritalStatus_Cerai	MaritalStatus_Duda	MaritalS
0	2	4	5	2	2	0	0	0	
1	2	4	4	0	0	0	0	0	
2	2	3	7	1	2	1	0	0	
3	2	1	2	0	0	1	0	0	
4	4	2	5	0	1	0	0	0	

Untuk selengkapnya, dapat melihat jupyter notebook disini

Preprocessing

Feature Scaling → Standardization

- Income
- Recency
- NumWebVisitsMonth
- Age
- TotalSpending
- NumOfTransactions

Before Scaling

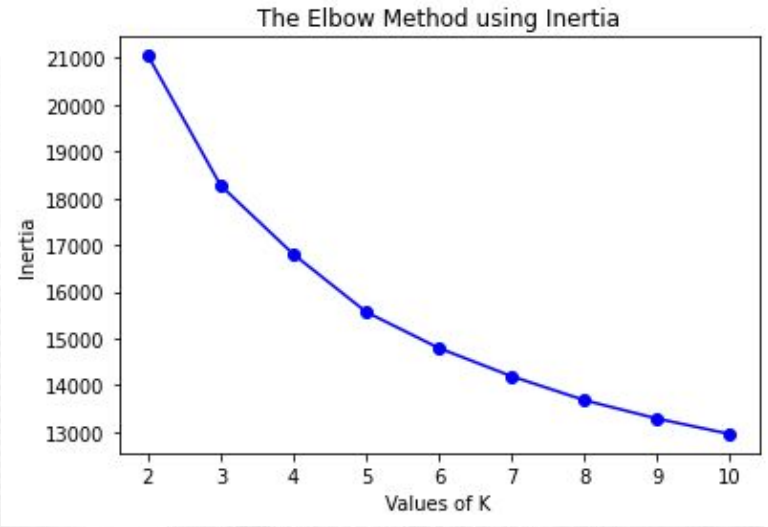
	Recency	NumWebVisitsMonth	Age	TotalSpent	NumOfTransactions
0	58	7	65	1617000	25
1	38	5	68	27000	6
2	26	4	57	776000	21
3	26	6	38	53000	8
4	94	5	41	422000	19

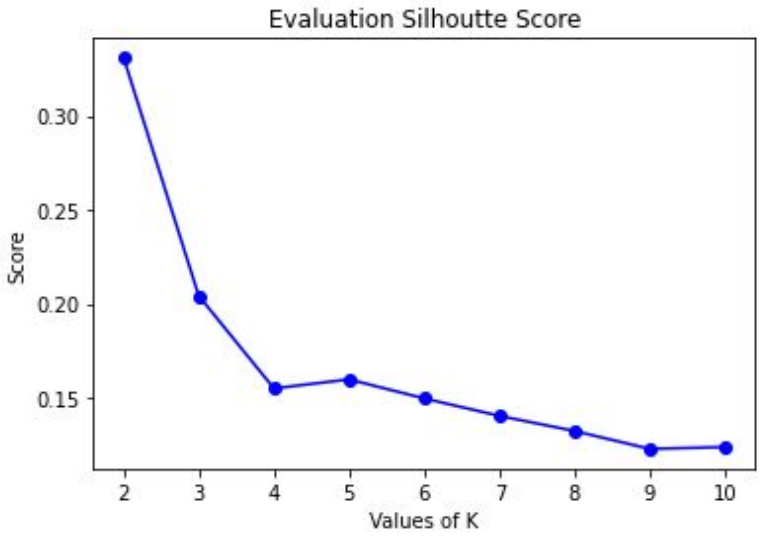
After Scaling

	Recency	NumWebVisitsMonth	Age	TotalSpent	NumOfTransactions
0	0.311959	0.686577	1.022565	1.684059	1.332699
1	-0.378363	-0.134431	1.280313	-0.961651	-1.162050
2	-0.792557	-0.544934	0.335236	0.284661	0.807488
3	-0.792557	0.276073	-1.297169	-0.918388	-0.899445
4	1.554540	-0.134431	-1.039421	-0.304384	0.544883

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling



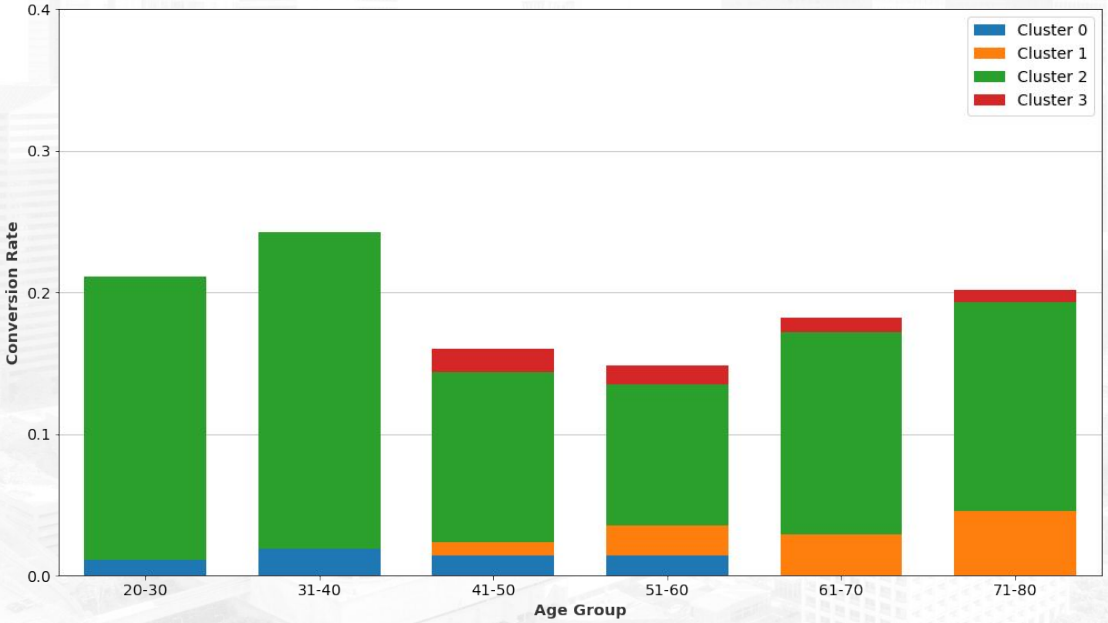


Berdasarkan hasil visualisasi **Elbow Method dan Silhoutte Score dapat disimpulkan bahwa jumlah cluster yang optimal adalah 4**. Pada Elbow Method, K=4 sudah mulai menunjukkan perubahan nilai inertia yang cukup konvergen untuk nilai K berikutnya. Sedangkan pada Silhoutte Score, K=4 merupakan score cukup baik jika dibandingkan dengan K lainnya

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

CVR based on Age in each Cluster



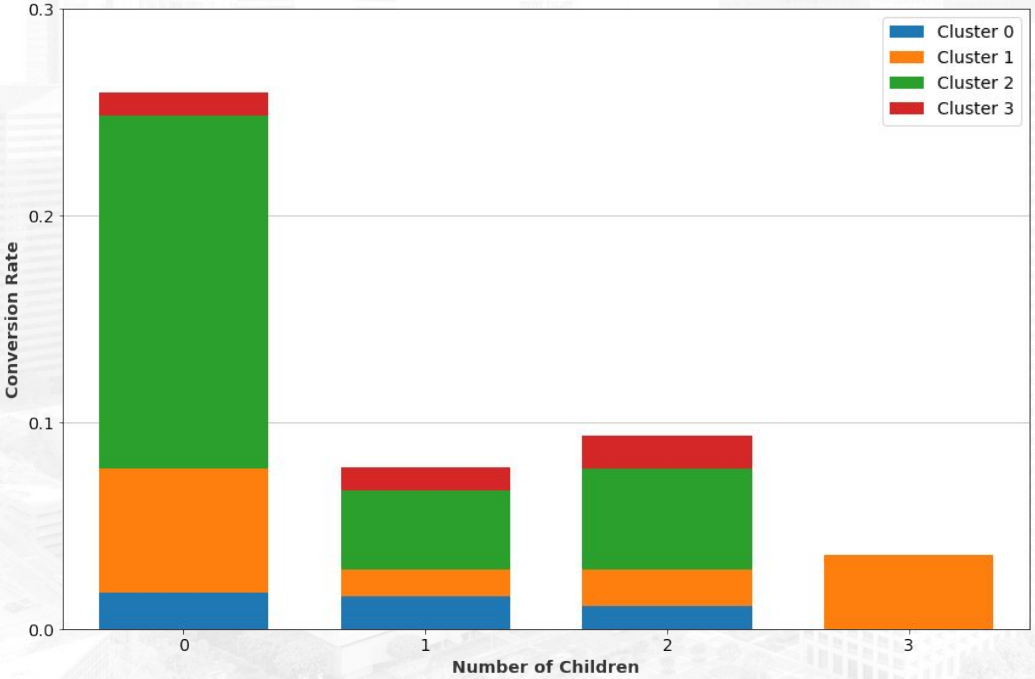
Insight:

- **Cluster 0** berusia 20-60 tahun dengan CVR yang kecil dan tidak mengalami perubahan.
- **Cluster 1** berusia 41-80 tahun dengan CVR yang semakin naik berdasarkan umurnya.
- **Cluster 2** berusia diatas 20 tahun dengan CVR yang paling diantara cluster lainnya.
- **Cluster 3** berusia 41-80 tahun dengan CVR yang paling kecil dan semakin menurun berdasarkan umur.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

CVR based on Children in each Cluster



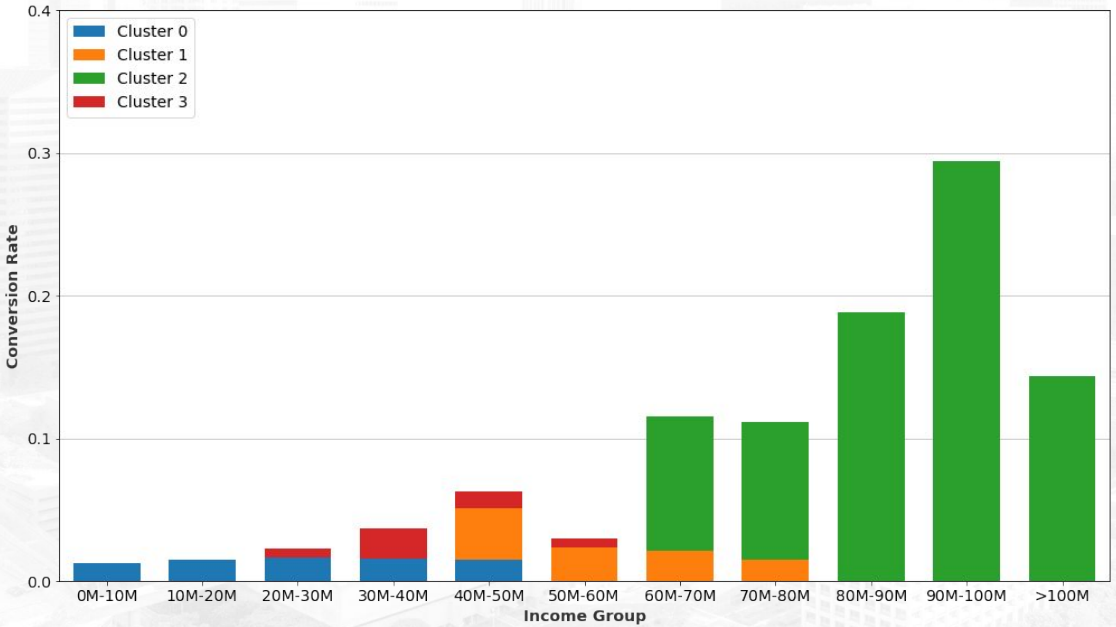
Insight:

- **Cluster 0** memiliki anak maksimal 2 dengan CVR yang tidak mengalami perubahan
- **Cluster 1** memiliki anak terutama 3 atau belum punya anak dengan CVR yang tinggi.
- **Cluster 2** lebih dominan pada pada yang belum memiliki anak karena memiliki CVR tertinggi dibandingkan lainnya.
- **Cluster 3** memiliki anak maksimal 3 dengan CVR yang tidak mengalami perubahan.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

CVR based on Income in each Cluster



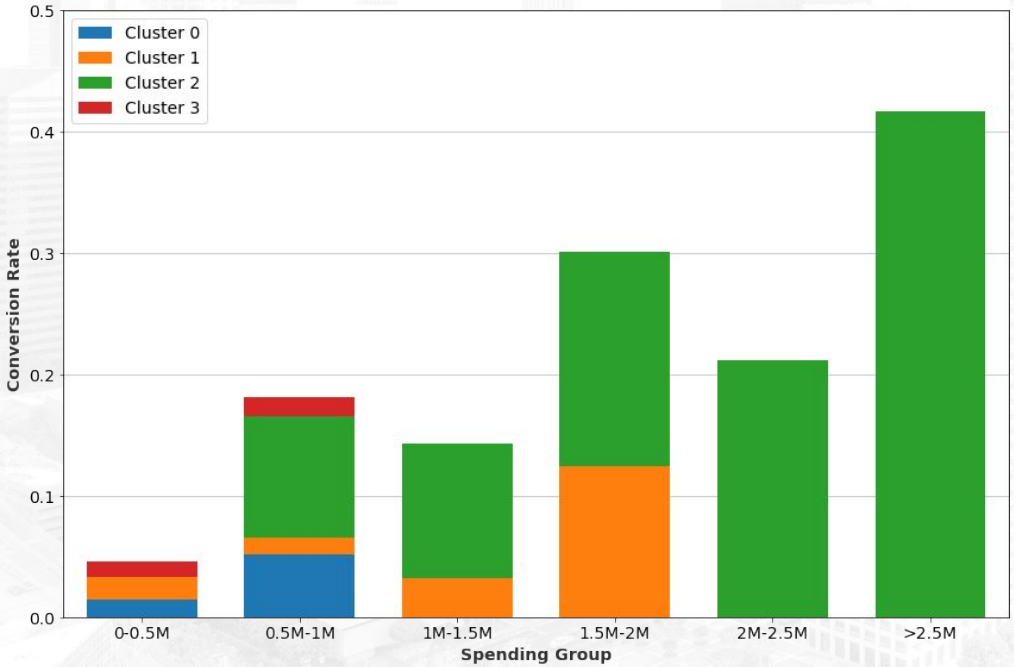
Insight:

- **Cluster 0** yang memiliki income dibawah 50 juta dengan CVR yang tidak mengalami perubahan.
- **Cluster 1** yang memiliki income 40-80 juta dengan CVR yang sedikit menurun tergantung incomenya.
- **Cluster 2** yang memiliki income diatas 60 juta dengan CVR tertinggi diantara cluster lainnya.
- **Cluster 3** yang memiliki income 20-60 juta dengan CVR yang tinggi pada kelompok income 30-40 juta.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

CVR based on Spending in each Cluster



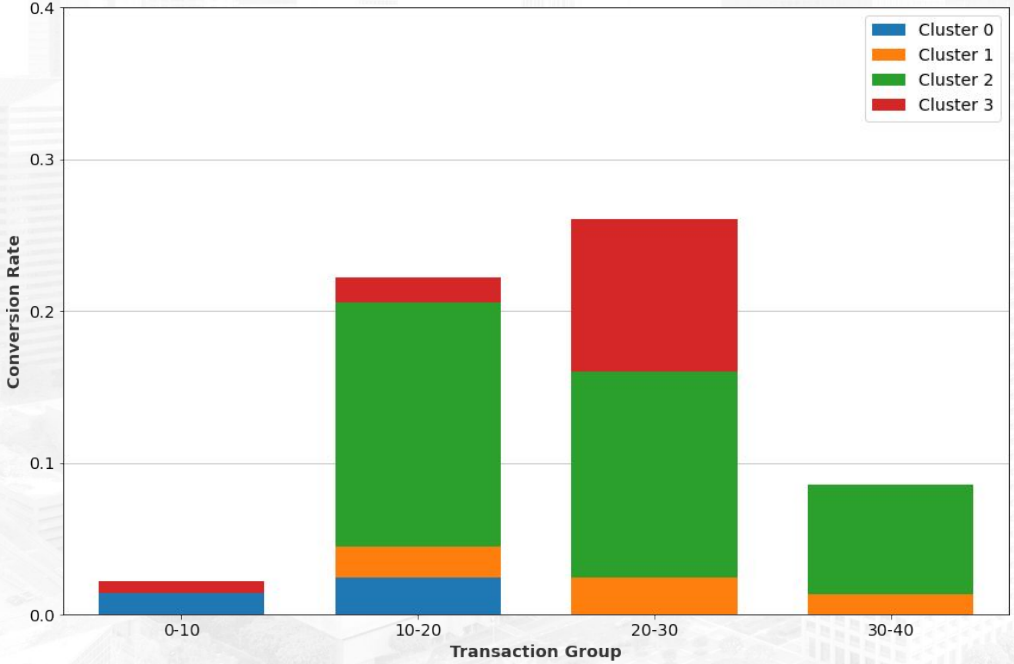
Insight:

- **Cluster 0** yang memiliki pengeluaran dibawah 1 juta dengan dengn CVR yang semakin naik tergantung pengeluarannya.
- **Cluster 1** yang memiliki pengeluaran dibawah 2 juta dengan CVR yang semakin naik tergantung pengeluarannya.
- **Cluster 2** yang memiliki pengeluaran diatas 500 ribu juta dengan CVR yang semakin naik tergantung pengeluarannya.
- **Cluster 3** yang memiliki pengeluaran dibawah 1 juta dengan CVR yang tidak mengalami perubahan.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

CVR based on Transaction in each Cluster



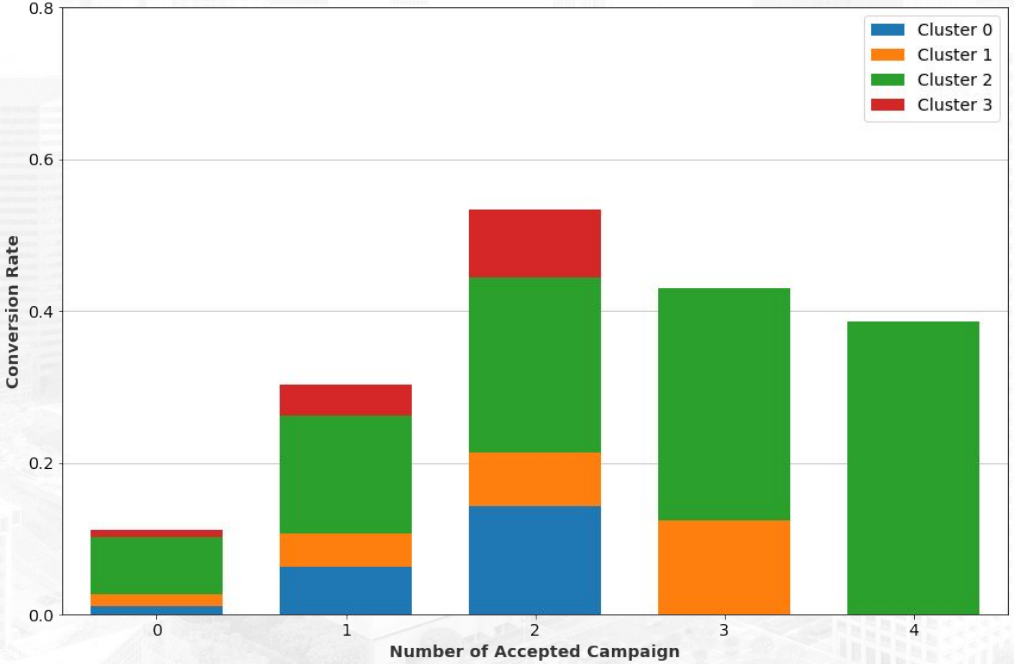
Insight:

- **Cluster 0** yang bertransaksi maksimal 20 kali dengan CVR yang hanya sedikit naik berdasarkan banyak transaksinya.
- **Cluster 1** yang bertransaksi 10-40 kali dengan CVR terendah yang hanya sedikit naik saat banyak transaksi 20-30 kali.
- **Cluster 2** yang bertransaksi 10-40 kali dengan CVR tertinggi tetapi mulai sedikit menurun seiring banyaknya transaksi.
- **Cluster 3** yang bertransaksi maksimal 30 kali dengan CVR yang semaik naik berdasarkan banyak transaksinya, terutam saat transaksi 20-30 kali.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

CVR based on Accepted Campaign in each Cluster



Insight:

- **Cluster 0** yang pernah menyetujui campaign maksimal 2 kali dengan CVR yang semakin naik tergantung banyaknya campaign.
- **Cluster 1** yang pernah menyetujui campaign maksimal 3 kali dengan CVR yang semakin naik tergantung banyaknya campaign.
- **Cluster 2** yang selalu menyetujui campaign yang diberikan dengan CVR tertinggi dan semakin naik tergantung banyaknya campaign.
- **Cluster 3** yang pernah menyetujui campaign maksimal 2 kali dengan CVR terendah tetapi semakin naik tergantung banyaknya campaign.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Conclusion

- **Cluster 0 (Low Spender)** adalah orang yang berusia 41-80 tahun, memiliki jumlah anak maksimal 2, mempunyai pendapatan diantara 20-60 juta, memiliki pengeluaran dibawah 1 juta, bertransaksi maksimal 20 kali, dan telah menyetujui jumlah campaign maksimal 2 kali.
- **Cluster 1 (Medium Spender)** adalah orang yang berusia 41-80 tahun, memiliki jumlah anak maksimal 3, mempunyai pendapatan diantara 40-80 juta, memiliki pengeluaran dibawah 2 juta, bertransaksi antara 10-40 kali, dan telah menyetujui jumlah campaign maksimal 3 kali.
- **Cluster 2 (High Spender)** adalah orang yang berusia diatas 20 tahun, memiliki jumlah anak maksimal 2, mempunyai pendapatan diatas 60 juta, memiliki pengeluaran diatas 500 ribu, bertransaksi diatas 10 kali, dan selalu menyetujui campaign yang diberikan.
- **Cluster 3 (Risk of Churn)** adalah orang yang berusia 20-60 tahun, memiliki jumlah anak maksimal 2, mempunyai pendapatan dibawah 50 juta, memiliki pengeluaran dibawah 1 juta, bertransaksi paling banyak 20-30 kali, dan telah menyetujui jumlah campaign maksimal 2 kali.

Recommendation

- **Penggunaan clustering dapat mempermudah retargeting marketing campaign** karena masing-masing cluster lebih terlihat jelas terutama dari faktor pendapatan dan pengeluaran yang berdampak posisif pada tingkat konversi.
- Untuk kelompok **High Spender** walaupun memiliki tingkat konversi tertinggi tetapi **harus tetap dilakukan treatment yang sama dengan kelompok lainnnya** agar revenue yang didapatkan tetap terjaga.
- Untuk kelompok **Medium Spender** hanya sedikit berbeda dengan High Spender dimana pendapatan lebih sedikit sehingga pengeluaran juga lebih sedikit. Namun transaksi yang dilakukan cukup sering, hal ini dimungkinkan transaksi yang dilakukan biasa menggunakan promo atau diskon sehingga **jumlah promo atau diskon yang dibuat perlu dibatasi**.
- Untuk kelompok **Low Spender** terlihat masih ragu dalam melakukan transaksi karena tingkat konversinya masih kecil walaupun transaksi yang dilakukan masih wajar mengingat jumlah pendapatan dan pengeluaran yang dimiliki. Kelompok ini perlu treatment tambahan agar menjadi lebih yakin dalam melakukan transaksi seperti **memberi promo atau diskon tambahan yang disesuaikan**.
- Untuk kelompok **Risk of Churn** adalah kelompok yang perlu diperhatikan karena tingkat konversi paling rendah dikarenakan pendapatan yang dihasilkan cukup paling kecil sehingga perlu dilakukan **treament ekstra seperti preferensi apa yang paling diminati** agar transaksi yang dilakukan naik tapi masih dalam kategori hemat **agar tidak beralih ke tempat lain**.

Rakamin Academy

Thank You !!

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