Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
Arieska Restu
arieskarestu02@gmail.com
linkedin.com/in/arieskarestu

I am an Assistant Lecturer who has experience in the field of Data Science with a background in Informatics. Experienced in Data Analysis, Data Mining, and Machine Learning projects. Also experienced in extracting primary and secondary data, as well as developing and maintaining databases. Able to conduct in-depth data analysis to identify trends that are relevant to companies and clients, and proficient in creating analysis reports. I also have expertise in programming languages and data visualization.

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 ”

Feature Engineering

- Conversion rate didapat dari hasil pembagian antara kolom Response dan NumWebVisitsMonth
- Age_Group didapatkan dari hasil perhitungan umur berdasarkan Year_Birth, kemudian dikelompokkan menjadi 5 kelompok. Kelompok tersebut yaitu Remaja, Dewasa Muda, Dewasa, Setengah Baya, dan Lanjut Usia.

Conversion Rate

ID	Response	NumWebVisitsMonth	Conversion_Rate
8041	0	9	0.0
5538	1	1	1.0
9381	0	3	0.0
3281	0	6	0.0
5067	0	5	0.0
5907	0	8	0.0
8690	0	2	0.0

Age Group

ID	Age	Age_Group
425	39	Dewasa
10240	75	Lanjut Usia
9888	55	Lanjut Usia
8754	50	Setengah Baya
1055	48	Setengah Baya
8876	61	Lanjut Usia
4643	51	Lanjut Usia

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Engineering

- Total_Kids didapat dari hasil penjumlahan anak berdasarkan Kidhome dan Teenhome.
- Total_Spending didapatkan dari hasil penjumlahan seluruh pengeluaran produk. Mulai dari MntCoke, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, hingga MntGoldProds.

Total Kids

ID	Kidhome	Teenhome	Total_Kids
10702	1	0	1
3068	0	0	0
5287	1	0	1
7660	1	1	2
6263	1	2	3
6927	1	1	2
3584	0	0	0

Total_Spending

ID	MntCoke	MntFruits	MntMeat Products	MntFish Products	MntSweet Products	MntGold Prods	Total_Spending
7476	928000	63000	254000	0	12000	12000	1269000
10951	279000	0	18000	0	0	9000	306000
10837	2000	1000	18000	20000	11000	16000	68000
2320	508000	11000	59000	23000	5000	29000	635000
1404	51000	23000	82000	33000	0	42000	231000
5067	966000	26000	282000	52000	26000	26000	1378000
4508	203000	35000	305000	46000	17000	227000	833000

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Engineering

- Total_Purchases didapatkan dari hasil penjumlahan seluruh jenis transaksi. Transaksi tersebut yaitu NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, dan NumStorePurchases.

Total_Purchases

ID	NumDeals Purchases	NumWeb Purchases	NumCatalog Purchases	NumStore Purchases	Total_Purchases
3068	1	2	0	3	6
1349	1	3	1	3	8
1183	1	1	0	2	4
5341	1	6	3	4	14
6825	1	1	0	2	4
10432	1	1	0	3	5
6379	3	2	2	2	9

Untuk selengkapnya, dapat melihat jupyter notebook disini

Exploration Data Analysis

- Proses EDA yang dilakukan terdiri dari tiga tahap, yakni mulai dari Quick EDA, Univariate Analysis, dan Multivariate Analysis.

- Proses-proses yang dilakukan pada tahap Quick EDA yaitu pengecekan informasi kolom dataset, pengecekan deskripsi statistic dari dataset, pengecekan data yang hilang, dan pengecekan data yang duplikat.

- Pada tahap Univariate Analysis dilakukan analisis dengan menggunakan visualisasi dari persebaran data untuk setiap kolom, baik kolom numerical maupun kolom categorical.

- Pada tahap Multivariate Analysis dilakukan analisis dengan menggunakan visualisasi dari correlation matrix untuk setiap fitur.

Untuk selengkapnya, dapat melihat jupyter notebook disini

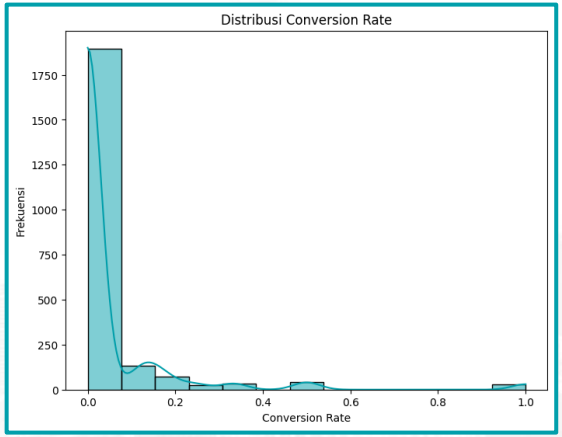
Exploration Data Analysis

- Dari proses EDA yang dilakukan didapatkan informasi-informasi dalam dataset, yakni sebagai berikut.
  - Terdapat fitur yang tidak memiliki nama.
  - Rata-rata dari conversion_rate cukup rendah yakni 0.04, hal ini menunjukkan bahwa customer yang mengunjungi website lalu melakukan transaksi sangat sedikit.
  - Distribusi dari total_spending tidak merata, hal ini terlihat dari mean dan median yang berbeda jauh.
  - Pada kolom Income terdapat 24 data yang missing value.
  - Pada kolom Conversion_Rate terdapat 11 data yang missing value, hal ini dikarenakan nilai pada kolom Response dan NumWebVisitsMonth adalah 0.
  - Dalam dataset tidak ada data yang duplicate.

- Selain itu, ditemukan beberapa fakta dan insight dari hasil analisis yang dilakukan dengan menggunakan dataset ini. Beberapa fakta dan insight tersebut yaitu
  - Sebagian besar conversion rate mendekati nol.
  - Sebagian besar pelanggan berada dalam rentang usia 40-60 tahun.
  - Kelompok dewasa muda adalah kelompok umur dengan pengeluaran tertinggi.
  - Kelompok dewasa muda adalah kelompok umur dengan total transaksi tertinggi.

Untuk selengkapnya, dapat melihat jupyter notebook disini

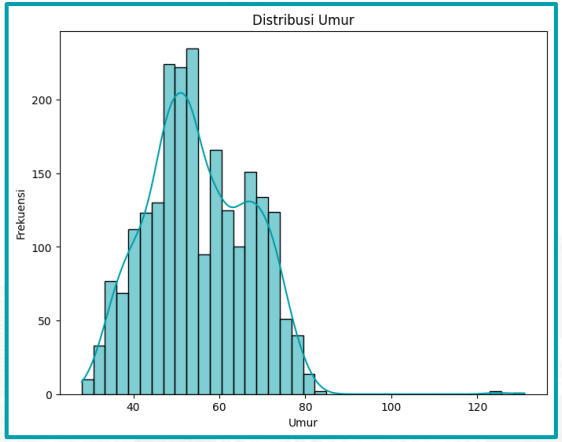
Fact 1: Sebagian besar conversion rate mendekati nol



Berdasarkan grafik tersebut, dapat diambil beberapa insight yaitu sebagai berikut.

- Efektivitas kampanye pemasaran yang kurang, mengingat sebagian besar Conversion Rate mendekati 0, kampanye pemasaran saat ini kurang efektif dalam mendorong pengunjung untuk berkonversi.
- Segmentasi pengunjung, mengingat ada sebagian kecil pengunjung dengan Conversion Rate yang tinggi, ini menunjukkan adanya segmen pengunjung yang merespon dengan baik. Perusahaan bisa fokus untuk mengidentifikasi karakteristik segmen ini dan menyesuaikan strategi pemasaran untuk menarik lebih banyak pengunjung seperti mereka.
- Perlu meningkatkan engagement, dengan banyaknya pengunjung yang memiliki Conversion Rate sangat rendah, perusahaan perlu mencari cara untuk meningkatkan engagement di website.

Fact 2: Mayoritas pelanggan di rentang usia 40-60 tahun



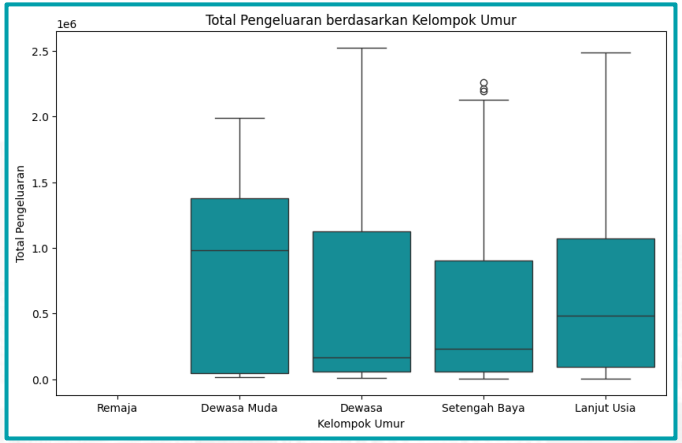
Dari grafik tersebut, dapat diambil beberapa insight yaitu sebagai berikut.

- Target pasar utama, dengan mayoritas pelanggan berada dalam rentang usia 40-60 tahun, perusahaan dapat menyimpulkan bahwa target pasar utama saat ini adalah kelompok usia menengah hingga lanjut usia. Oleh karena itu, strategi pemasaran dan produk yang dikembangkan harus relevan dengan preferensi dan kebutuhan mereka.

- Peluang untuk segmentasi pasar, mengingat adanya variasi dalam distribusi umur, perusahaan memiliki peluang untuk melakukan segmentasi pasar lebih lanjut berdasarkan usia. Misalnya, perusahaan bisa menciptakan kampanye pemasaran yang lebih khusus dan relevan untuk kelompok usia yang lebih muda (30-40 tahun) atau lebih tua (60 tahun ke atas).

- Adanya outlier di usia lebih dari 100 tahun, menunjukkan bahwa adanya segmen pelanggan yang sangat tua.

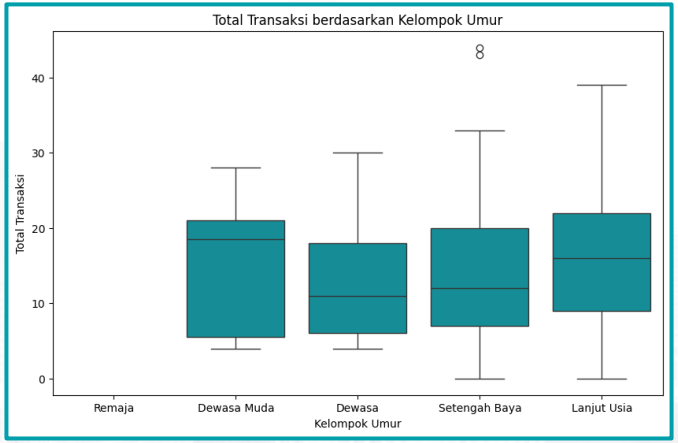
Fact 3: Dewasa muda adalah kelompok pengeluaran tertinggi



Dari grafik tersebut, dapat diambil beberapa insight yaitu sebagai berikut.

- Kelompok umur dengan pengeluaran tertinggi, yakni kelompok dewasa muda bisa menjadi target pasar yang menarik untuk kampanye pemasaran yang difokuskan pada produk atau layanan premium.
- Preferensi belanja berdasarkan kelompok umur, hal ini bisa menunjukkan preferensi belanja yang berbeda. Dengan memahami preferensi ini bisa membantu perusahaan menyesuaikan penawaran produk mereka.
- Analisis outlier untuk strategi khusus, pelanggan yang termasuk outliers mungkin merupakan high spenders yang bisa diberikan penawaran khusus atau layanan premium untuk meningkatkan loyalitas. Di sisi lain, outliers dengan pengeluaran sangat rendah juga dapat dianalisis untuk memahami hambatan pengeluaran mereka dan mengembangkan strategi untuk mengatasi hal ini.

Fact 3: Dewasa muda adalah kelompok total transaksi tertinggi



Dari grafik tersebut, dapat diambil beberapa insight yaitu sebagai berikut.

- Kelompok umur dengan aktivitas transaksi tertinggi, yakni kelompok dewasa muda dimana kelompok ini lebih aktif berbelanja. Perusahaan dapat memfokuskan upaya pemasaran mereka untuk meningkatkan engagement dengan kelompok ini, misalnya melalui penawaran khusus atau program loyalitas.
- Variasi pola perilaku transaksi berdasarkan kelompok umur, menunjukkan perbedaan perilaku pembelian. Kelompok yang lebih muda mungkin lebih sering melakukan pembelian dalam jumlah kecil, sementara kelompok yang lebih tua mungkin melakukan pembelian yang lebih jarang namun dalam jumlah yang lebih besar.
- Analisis outlier untuk penawaran khusus, outliers dengan jumlah transaksi yang sangat tinggi bisa jadi merupakan pelanggan setia atau pelanggan dengan potensi tinggi. Perusahaan dapat mempertimbangkan untuk memberikan penawaran khusus atau penghargaan kepada pelanggan ini untuk meningkatkan loyalitas mereka. Sebaliknya, outliers dengan jumlah transaksi yang sangat rendah mungkin memerlukan pendekatan yang berbeda untuk meningkatkan keterlibatan mereka.

Data Cleaning

- Berdasarkan hasil dari tahap Exploratory Data Analysis, terdapat data null yakni pada kolom Income sebanyak 24 data dan Conversion_Rate sebanyak 11 data. Selain itu, tidak terdapat data yang duplicate.

Conversion Rate

ID	Response	Conversion_Rate	NumWebVisitsMonth
8475	0	NaN	0
5555	0	NaN	0
1501	0	NaN	0
11074	0	NaN	0
10286	0	NaN	0
8584	0	NaN	0
6237	0	NaN	0

Income

ID	Year_Birth	Education	Marital_Status	Income
1994	1983	S1	Menikah	NaN
5255	1986	S1	Lajang	NaN
7281	1959	S3	Lajang	NaN
7244	1951	S1	Lajang	NaN
8557	1982	S1	Lajang	NaN
10629	1973	D3	Menikah	NaN
8996	1957	S3	Menikah	NaN

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning

- Penanganan data null yaitu dilakukan dengan mengisi nilainya dengan mean untuk fitur Income dan nilai 0 untuk fitur Conversion_Rate.

Conversion Rate

ID	Response	Conversion_Rate	NumWebVisitsMonth
8475	0	0	0
5555	0	0	0
1501	0	0	0
11074	0	0	0
10286	0	0	0
8584	0	0	0
6237	0	0	0

Income

ID	Year_Birth	Education	Marital_Status	Income
1994	1983	S1	Menikah	52247251.35
5255	1986	S1	Lajang	52247251.35
7281	1959	S3	Lajang	52247251.35
7244	1951	S1	Lajang	52247251.35
8557	1982	S1	Lajang	52247251.35
10629	1973	D3	Menikah	52247251.35
8996	1957	S3	Menikah	52247251.35

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

- Pada tahap data preprocessing, dilakukan beberapa proses yakni proses drop features, feature encoding, dan standardization.

- Untuk proses drop features dilakukan untuk menghapus fitur-fitur yang tidak diperlukan. Dimana fitur-fitur yang dihapus yakni fitur-fitur yang telah diolah pada tahap feature engineering dan fitur yang tidak diperlukan seperti 'Unnamed: 0', 'ID', ‘Dt_Customer’, 'Z_CostContact', dan 'Z_Revenue’

- Setelah dilakukan proses drop features, menghasilkan kolom sebanyak 16 kolom.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

- Pada proses feature encoding, melakukan proses encoding untuk fitur-fitur categorical yakni 'Education', 'Marital_Status', dan 'Age_Group'.

- Untuk urutan dari fitur Education dari yang terendah ke tertinggi yakni SMA, D3, S1, S2, S3.

- Untuk urutan dari fitur Marital_Statusdari yang terendah ke tertinggi yakni 'Lajang’, 'Bertunangan', 'Menikah', 'Cerai', 'Janda', 'Duda’.

- Untuk urutan dari fitur Age_Group dari yang terendah ke tertinggi yakni 'Remaja', 'Dewasa Muda', 'Dewasa', 'Setengah Baya', 'Lanjut Usia'.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

- Pada proses feature encoding ini dilakukan dengan menggunakan metode label encoding.

Fitur categorical (*Before*)

Education	Marital_Status	Age_Group
S1	Lajang	Lanjut Usia
S1	Lajang	Lanjut Usia
S1	Bertunangan	Lanjut Usia
S1	Bertunangan	Dewasa
S3	Menikah	Setengah Baya
S2	Bertunangan	Lanjut Usia
S1	Cerai	Lanjut Usia

Fitur categorical (*After*)

Education	Marital_Status	Age_Group
2	0	4
2	0	4
2	1	4
2	1	2
4	2	3
3	1	4
2	3	4

Untuk selengkapnya, dapat melihat jupyter notebook disini

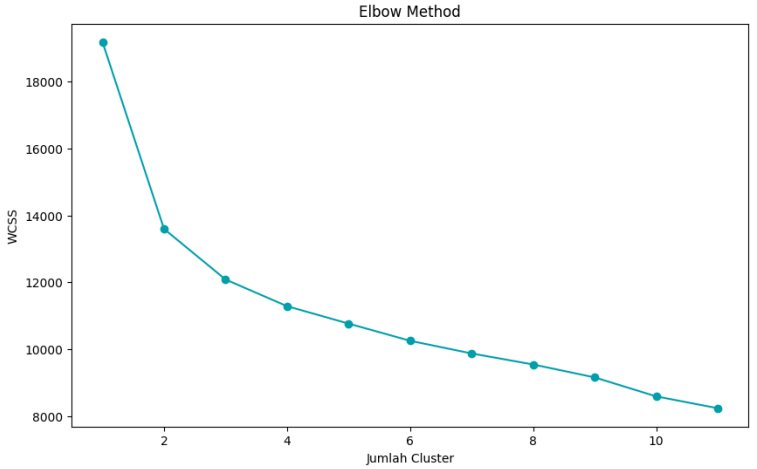
Data Preprocessing

- Setelah melakukan proses drop features dan feature encoding, proses selanjutnya yaitu standardization.
- Proses standardization dilakukan untuk setiap fitur numerik agar berada ke skala di mana nilai rata-rata 0 dan standar deviasi adalah 1.
- Fitur-fitur yang dilakukan standarisasi yakni 'Income', 'Recency', 'Total_Kids', 'Total_Spending', dan 'Total_Purchases'.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling

- Proses yang pertama dilakukan yakni menentukan jumlah cluster dengan Elbow Method. Berikut hasil visualisasinya.



- Dari grafik tersebut, titik siku berada di sekitar cluster ke-5. Setelah titik ini, penurunan WCSS menjadi lebih landai dan tidak terlalu signifikan. Oleh karena itu, jumlah cluster yang tepat untuk digunakan adalah 5 cluster.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling

- Setelah menentukan jumlah cluster, selanjutnya yaitu melakukan implementasi clustering dengan menggunakan K-Means Clustering.
- Dimana jumlah cluster yang dipakai adalah sebesar 5 cluster.
- Setelah itu, melakukan evaluasi dengan menggunakan Silhouette Score.
- Hasil Silhouette Score yakni sebesar 0,2585.
- Nilai 0,2585 lebih dekat ke 0 daripada 1, yang menunjukkan bahwa clustering yang dilakukan oleh K-Means dengan 5 cluster tidak terlalu baik. Data tidak dikelompokkan secara jelas atau ada beberapa titik data yang tidak sesuai dengan cluster mereka.

Untuk selengkapnya, dapat melihat jupyter notebook disini

THANK YOU

Have a nice day!