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:
Muhammad Cikal Merdeka
Email : mcikalmerdeka@gmail.com
LinkedIn : linkedin.com/in/mcikalmerdeka
Github : github.com/mcikalmerdeka

Dedicated entry-level data scientist with analytical and experimental background of Physics. My graduation 2023, a pivotal year marked by significant advancements in artificial intelligence with the introduction of GPT-4 and other generative AI models, has fueled my curiosity and excitement to delve into the field of data. I have comprehensive grasp of data science methodology from business understanding to modelling process with proficiency in **Python, SQL, Tableau, Power BI, Looker Studio and other tools** related to data analytics workflow from several coursework and bootcamps.

Data Cleaning & Preprocessing

Rakamin Academy

Drop Unnecessary/Unrelated Columns

- Some columns (not all) in the dataset that have lost their importance as an individual column will be dropped since their information already stored in the new engineered column (as explained in Task 1).

- These columns also doesn't store related information to our model and keeping them will just increase the dimension.

- Columns that will be dropped are :
  - ❑ Unnamed: 0
  - ❑ ID
  - ❑ Year_Birth
  - ❑ Marital_Status
  - ❑ Dt_Customer

For more details, you can view the Jupyter notebook here.

Rakamin Academy

Identifying Missing and Duplicated Values

	Feature	Data Type	Null Values	Null Percentage	Duplicated Values
0	ID	int64	0	0.00	0
1	Year_Birth	int64	0	0.00	0
2	Education	object	0	0.00	0
3	Marital_Status	object	0	0.00	0
4	Income	float64	24	1.07	0
5	Kidhome	int64	0	0.00	0
6	Teenhome	int64	0	0.00	0
7	Dt_Customer	object	0	0.00	0
8	Recency	int64	0	0.00	0
9	MntCoke	int64	0	0.00	0
10	MntFruits	int64	0	0.00	0
11	MntMeatProducts	int64	0	0.00	0
12	MntFishProducts	int64	0	0.00	0
13	MntSweetProducts	int64	0	0.00	0
14	MntGoldProds	int64	0	0.00	0
15	NumDealsPurchases	int64	0	0.00	0
16	NumWebPurchases	int64	0	0.00	0
17	NumCatalogPurchases	int64	0	0.00	0
18	NumStorePurchases	int64	0	0.00	0
19	NumWebVisitsMonth	int64	0	0.00	0
20	AcceptedCmp3	int64	0	0.00	0
21	AcceptedCmp4	int64	0	0.00	0
22	AcceptedCmp5	int64	0	0.00	0
23	AcceptedCmp1	int64	0	0.00	0
24	AcceptedCmp2	int64	0	0.00	0
25	Complain	int64	0	0.00	0
26	Z_CostContact	int64	0	0.00	0
27	Z_Revenue	int64	0	0.00	0
28	Response	int64	0	0.00	0

```python
df['Income'] = df['Income'].fillna(df['Income'].median())
```

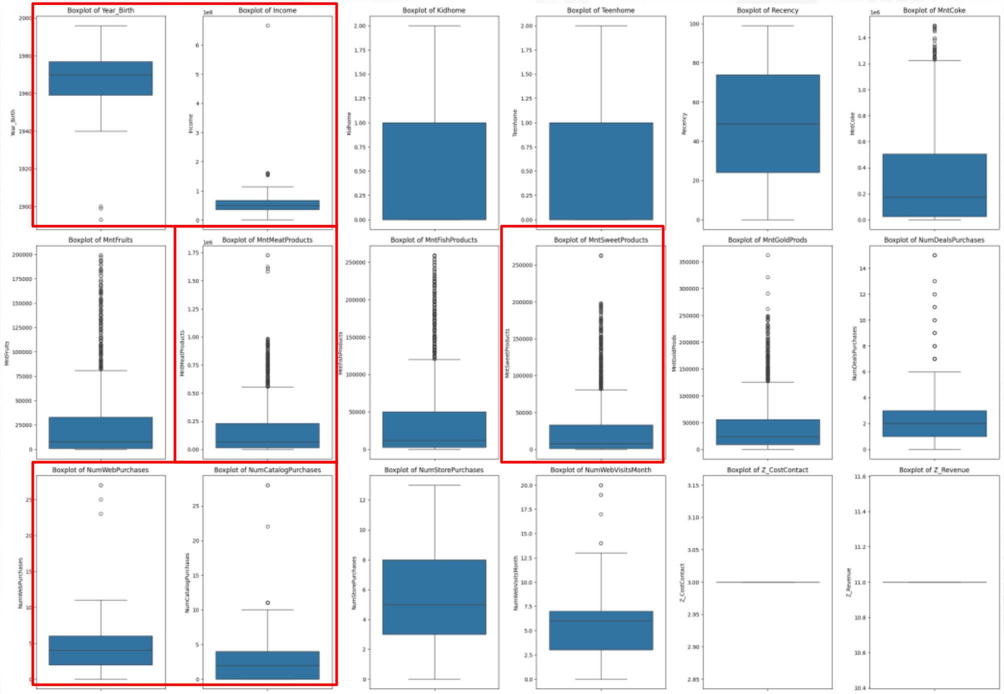
- Income column have a small percentage missing values, we will handle that by **imputation with median values** considering the positive skewed distribution of the values.
- No duplicated values in this dataset.

For more details, you can view the Jupyter notebook here.

Data Cleaning & Preprocessing

Outliers Handling

Handling outliers with z-score method to columns that have really extreme values (not applied to all columns). Columns that will be handled are : Year_Birth, Income, MntMeatProducts, MntSweetProducts, NumWebPurchases, NumCatalogPurchases



- Rows before removing outliers: 2240
- Rows after removing outliers: 1861

Feature Encoding

There are 3 categorical column that will be encoded based on their data type.

- Education : Ordinal type → Label encoding
- Age_Group : Ordinal type → Label encoding
- Marital_Status : Nominal type → One-hot encoding

	Education	Age_Group	Marital_Status
1	S1	Senior Adult	Lajang
2	S1	Middle Adult	Bertunangan
3	S1	Middle Adult	Bertunangan
4	S3	Middle Adult	Menikah
5	S2	Middle Adult	Bertunangan

Before Encoding

	Education	Age_Group	Marital_Status_Bertunangan	Marital_Status_Cerai	Marital_Status_Duda	Marital_Status_Janda	Marital_Status_Lajang	Marital_Status_Menikah
1	2	2	0	0	0	0	1	0
2	2	1	1	0	0	0	0	0
3	2	1	1	0	0	0	0	0
4	4	1	0	0	0	0	0	1
5	3	1	1	0	0	0	0	0

After Encoding

Rakamin Academy

Feature Scaling

- After we have all the values in numerical form, we need to transform (scale) the values to ensure fair calculations, especially when we will use distance-based algorithms like K-means clustering.
- We use standardization for scaling method to ensure that all columns have mean of 0 and standard deviation of 1.

```python
# Standardization
from sklearn.preprocessing import StandardScaler
ss = StandardScaler()
df_std_values = ss.fit_transform(df_preprocessed[df_preprocessed.columns])
```

	mean	std
Education	-1.240873e-16	1.000269
Income	-1.756312e-16	1.000269
Kidhome	1.088150e-16	1.000269
Teenhome	0.000000e+00	1.000269
Recency	1.221782e-16	1.000269
MntCoke	-4.199876e-17	1.000269
MntFruits	-1.336324e-17	1.000269
MntMeatProducts	-3.818070e-18	1.000269
MntFishProducts	-7.636139e-18	1.000269
MntSweetProducts	-5.345297e-17	1.000269
MntGoldProds	4.199876e-17	1.000269
NumDealsPurchases	-1.450866e-16	1.000269
NumWebPurchases	1.336324e-17	1.000269
NumCatalogPurchases	7.636139e-17	1.000269
NumStorePurchases	1.909035e-18	1.000269
NumWebVisitsMonth	1.679951e-16	1.000269
AcceptedCmp3	-2.863552e-17	1.000269
AcceptedCmp4	8.781560e-17	1.000269
AcceptedCmp5	-3.054456e-17	1.000269
AcceptedCmp1	7.636139e-17	1.000269
AcceptedCmp2	9.545174e-18	1.000269
Complain	4.056699e-17	1.000269
Z_CostContact	0.000000e+00	0.000000
Z_Revenue	0.000000e+00	0.000000
Response	3.054456e-17	1.000269
Age	-1.622680e-16	1.000269
Age_Group	0.000000e+00	1.000269
Num_Child	-1.489047e-16	1.000269
Membership_Duration	7.788862e-16	1.000269
Total_Acc_Camp	7.492961e-17	1.000269
Total_Spending	3.818070e-18	1.000269
Total_Purchases	4.390780e-17	1.000269
CVR	-8.017946e-17	1.000269
Marital_Status_Bertunangan	7.874768e-17	1.000269
Marital_Status_Cerai	-3.197633e-17	1.000269
Marital_Status_Duda	1.909035e-17	1.000269
Marital_Status_Janda	-4.486232e-17	1.000269
Marital_Status_Lajang	4.868039e-17	1.000269
Marital_Status_Menikah	-5.727104e-18	1.000269

After Scaling