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.

Initial Transformation with PCA Method (Optional)

Before we evaluate how many cluster should we make, in this case since we still have a large number of features because we didn't do the feature selection process, we need to reduce the amount of features into 2 with PCA first.

```python
from sklearn.decomposition import PCA

# Fit pca
pca = PCA(n_components = 2)
pca.fit(df_model)

# Transform data
data_pca = pca.transform(df_model)
df_pca = pd.DataFrame(data_pca, columns = ['PC 1', 'PC 2'])
```

PCA Implementation

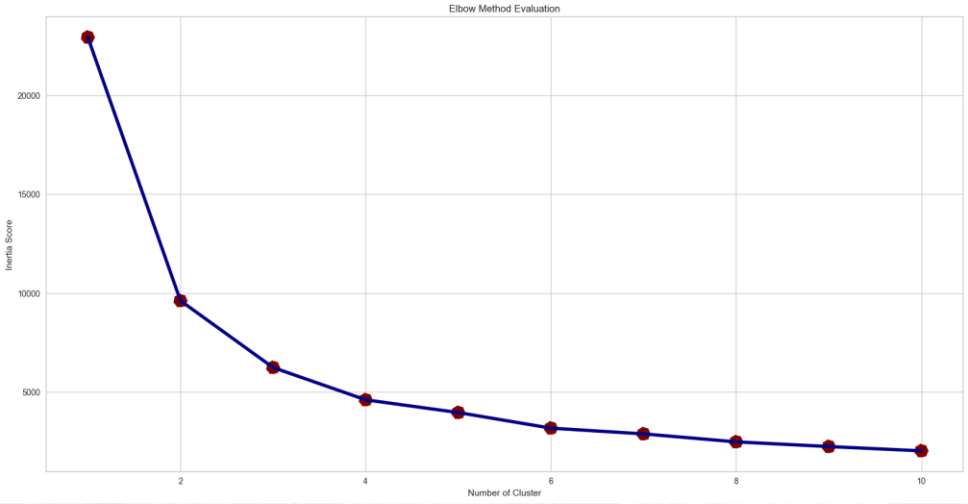
Dataframe After PCA

The preprocessed dataframe that contains 39 features is condensed into 2 features.

	PC 1	PC 2
0	-2.587152	-0.998152
1	3.024975	1.310848
2	-2.761262	1.390182
3	0.644242	0.418708
4	2.154105	-0.732064
...	...	...
1856	-3.164284	1.251221
1857	0.425386	-4.400500
1858	3.170721	1.312265
1859	3.687071	-0.933770
1860	-1.314482	-2.478322

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

Elbow Method

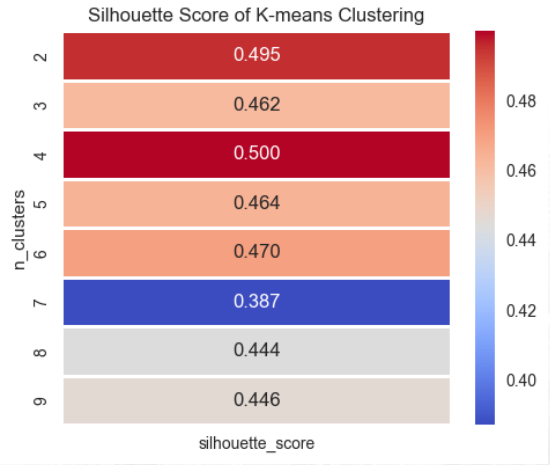


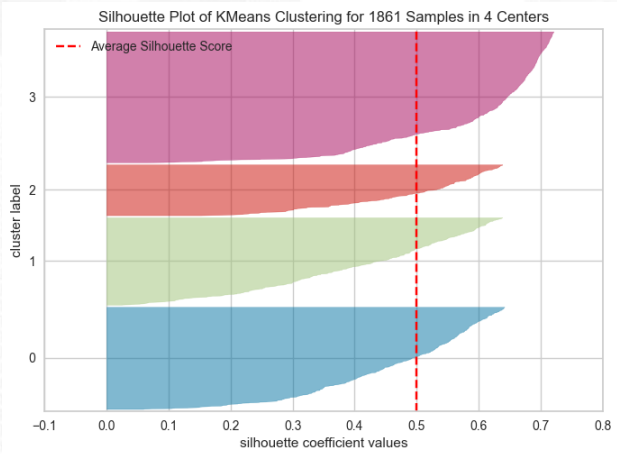
0	22916.944584
1	9615.884407
2	6243.349675
3	4611.780034
4	3969.076507
5	3182.426916
6	2892.136420
7	2489.284382
8	2253.162567
9	2036.257323

- Evaluating the elbow method we could see that **the optimal number of cluster is 4**. This is obtained by observing the inertia score, which begins to show similarity in subsequent clusters starting from the 4th cluster.

- Inertia is a metric that measures the sum of squared distances between each data point and its nearest centroid. It reflects how compact the clusters are, with lower values indicating tighter, more cohesive clusters.

Silhouette Score





- Based on the Silhouette Score, **the recommended optimal number of clusters is 4**.
- The Silhouette Score for this number of clusters is higher compared to other numbers of clusters, indicating better clustering quality.
- Silhouette Score is an evaluation metric that describes how well objects within one cluster are grouped within their own data compared to other clusters. The higher the Silhouette Score, the better the separation of clusters.