Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
atinazr@gmail.com
https://www.linkedin.com/in/atinazr/

“Atina is a junior data scientist with a bachelor’s degree in applied computer technology. Experienced in performing data analyst, visualization, and model building.”

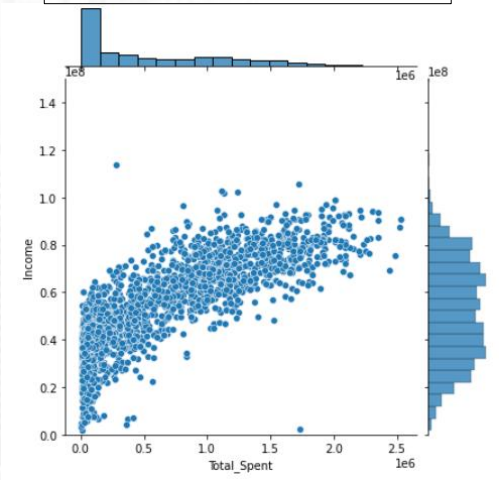
Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

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 ”

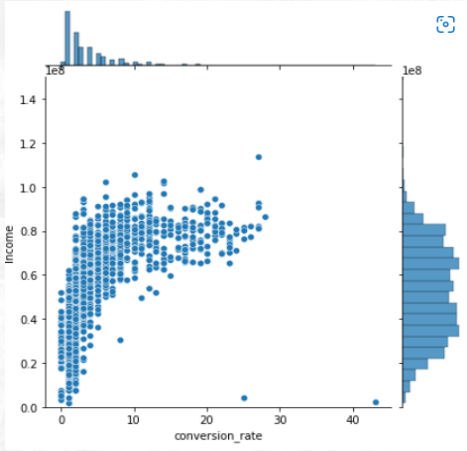
Conversion Rate Analysis Based on Income, Spending and Age

Income terhadap total pengeluaran



Fitur income dan fitur total_spent menunjukkan hubungan yang positif. Peningkatan yang terjadi pada fitur income juga diikuti peningkatan pada fitur total_spent.

Income terhadap total Conversion Rate

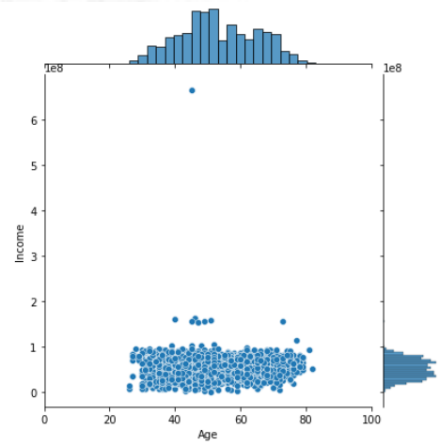


Ada kecenderungan korelasi positif, jika x naik y cenderung naik, namun ada beberapa factor lain yang turut memengaruhi.

Untuk selengkapnya, dapat melihat jupyter notebook disini

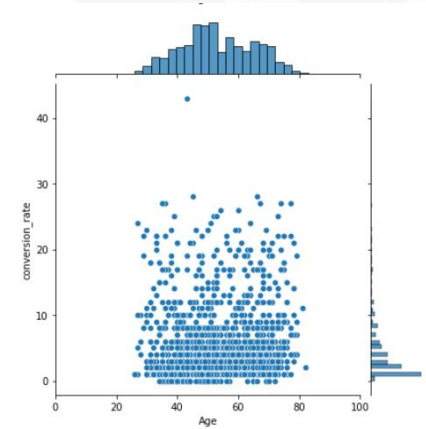
Conversion Rate Analysis Based on Income, Spending and Age

Income terhadap Umur



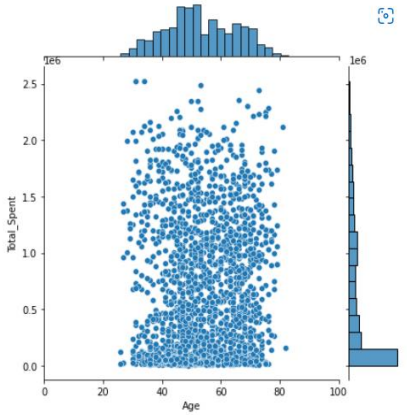
Data cenderung berkumpul, dimana tidak ada korelasi kuat antara income terhadap umur.

Conversion Rate terhadap umur



Tidak terlihat adanya korelasi antara conversion rate dan umur.

Pengeluaran terhadap Umur



Tidak terlihat adanya korelasi antara total pengeluaran dan umur.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning & Preprocessing

Rakamin Academy

Missing Value :

```
df.isna().sum()
```

```
ID                     0
Year_Birth             0
Education              0
Marital_Status         0
Income                24
Kidhome                0
Teenhome               0
Dt_Customer            0
Recency                0
MntCoke                0
MntFruits              0
MntMeatProducts        0
MntFishProducts        0
MntSweetProducts       0
MntGoldProds           0
NumDealsPurchases      0
NumWebPurchases        0
NumCatalogPurchases    0
NumStorePurchases      0
NumWebVisitsMonth      0
AcceptedCmp3           0
AcceptedCmp4           0
AcceptedCmp5           0
AcceptedCmp1           0
AcceptedCmp2           0
Complain               0
Z_CostContact          0
Z_Revenue              0
Response               0
dtype: int64
```

Check Duplicate :

```
print(df.duplicated().sum())
df.shape
```

```
0

(2216, 29)
```

Handle Missing Value :

Delete 24 Rows of missing value. The number is only little bit, so it's safe to delete.

```
df = df.dropna()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2216 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2216 non-null   int64
 1   Year_Birth           2216 non-null   int64
 2   Education            2216 non-null   object
 3   Marital_Status       2216 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2216 non-null   int64
 6   Teenhome             2216 non-null   int64
 7   Dt_Customer          2216 non-null   object
 8   Recency              2216 non-null   int64
 9   MntCoke              2216 non-null   int64
 10  MntFruits            2216 non-null   int64
 11  MntMeatProducts      2216 non-null   int64
 12  MntFishProducts      2216 non-null   int64
 13  MntSweetProducts     2216 non-null   int64
 14  MntGoldProds         2216 non-null   int64
 15  NumDealsPurchases    2216 non-null   int64
 16  NumWebPurchases      2216 non-null   int64
 17  NumCatalogPurchases  2216 non-null   int64
 18  NumStorePurchases    2216 non-null   int64
 19  NumWebVisitsMonth    2216 non-null   int64
 20  AcceptedCmp3         2216 non-null   int64
 21  AcceptedCmp4         2216 non-null   int64
 22  AcceptedCmp5         2216 non-null   int64
 23  AcceptedCmp1         2216 non-null   int64
 24  AcceptedCmp2         2216 non-null   int64
 25  Complain             2216 non-null   int64
 26  Z_CostContact        2216 non-null   int64
 27  Z_Revenue            2216 non-null   int64
 28  Response             2216 non-null   int64
dtypes: float64(1), int64(25), object(3)
memory usage: 519.4+ KB
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Encoding

1. Label Encoder

```
# label encoder
data = df.copy()
map_education = {
    'SMA' : 0,
    'D3' : 1,
    'S1' : 2,
    'S2' : 3,
    'S3' : 4
}

data['Education'] = data['Education'].map(map_education)
```

```
data[['Education']].sample(5)
```

	Education
1115	2
1590	4
593	2
2077	0
1821	3

2. One Hote Encoder

```
#One hot encoder
cat_ohe = ['Marital_Status', 'Age_Group']
data = pd.get_dummies(data=data, columns=cat_ohe)
```

```
36  Marital_Status_Bertunangan   2216 non-null   uint8
37  Marital_Status_Cerai         2216 non-null   uint8
38  Marital_Status_Duda          2216 non-null   uint8
39  Marital_Status_Janda         2216 non-null   uint8
40  Marital_Status_Lajang        2216 non-null   uint8
41  Marital_Status_Menikah       2216 non-null   uint8
42  Age_Group_mature             2216 non-null   uint8
43  Age_Group_middle aged        2216 non-null   uint8
44  Age_Group_old                2216 non-null   uint8
45  Age_Group_young              2216 non-null   uint8
```

Untuk selengkapnya, dapat melihat jupyter notebook disini

STANDARDIZATION

- Standarisasi dilakukan untuk merubah bentuk sebaran data menjadi mendekati distribusi normal

```
nums = ['Income', 'Kidhome', 'Teenhome', 'Recency', 'MntCoke', 'MntFruits', 'MntMeatProducts', 'MntFishProducts',
        'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases','NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases',
        'NumWebVisitsMonth', 'Z_CostContact', 'Z_Revenue','Age', 'Total_Kids', 'Total_Spent', 'Total_Transact']
```

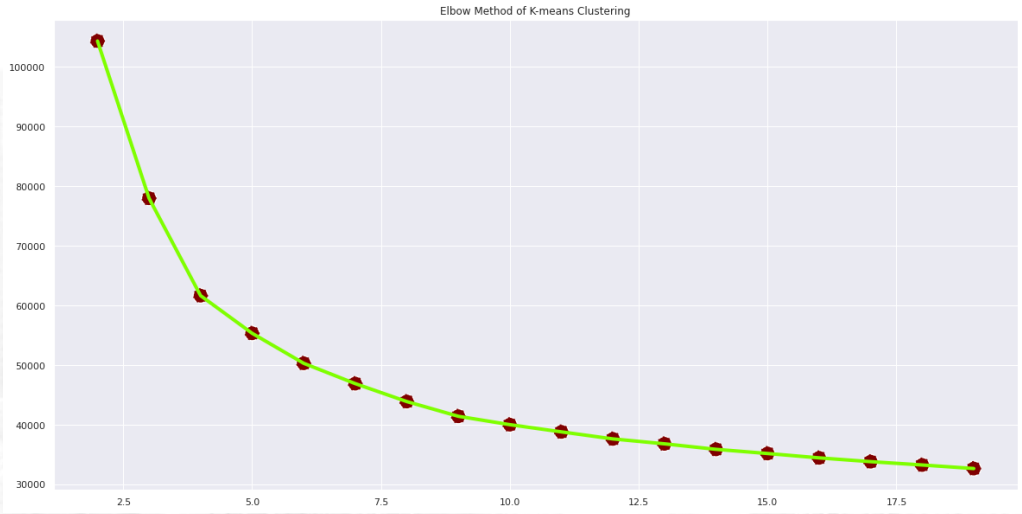
```
from sklearn.preprocessing import StandardScaler

for col in nums:
    data[col] = StandardScaler().fit_transform(data[[col]])

display(data.shape, data.head(5))
```

```
(2216, 46)
```

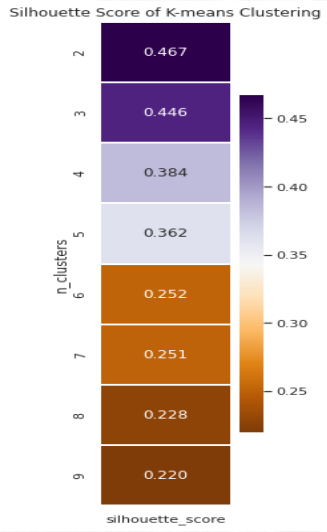
Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling

Visualisasi Elbow Method



Silhouette Score



Kedua visualisasi di atas digunakan untuk melihat jumlah cluster yang didapatkan dari penggunaan algoritma K-means clustering.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Customer Personality Analysis for Marketing Retargeting

INTERPRETASI :

1. Low Spender:
- Didominasi oleh kelompok usia >55 tahun dan 36-55 tahun
- cukup sering mencari promo yang masing-masing orangnya membeli promo 2 kali dalam sebulan (median)
- mempunyai total pendapatan dan pengeluaran yang kecil

2. Risk of Churn:
- Kelompok dengan jumlah user terbesar sebanyak 900 orang yang di dominasi oleh usia 36-55 tahun
- Mempunyai pendapatan dan pengeluaran paling kecil di setiap bulannya.
- Kelompok yang paling sering mengunjungi web dengan median total kunjungan 7 kali dalam sebulan. Namun, jarang melakukan transaksi
- Tidak banyak merespon campaign

3. Mid Spender:
- Didominasi oleh usia >55 tahun dan 36-55 tahun
- Memmpunyai total pendapatan dan pengeluaran terbesar kedua dibandingkan Kelompok lainnya
- Walaupun cukup jarang untuk visit web, namun paling sering merespon campaign dan menggunakan promo dalam sebulannya

4. High Spender:
- Kelompok dengan jumlah user terkecil sebanyak 137 orang yang di dominasi oleh usia >55 tahun) dan 36-55 tahun
- Mempunyai pendapatan dan pengeluaran paling besar di setiap bulannya
- Jumlah penggunaan promo paling sedikit dibandingkan dengan yang lainnya.
- Mempunyai conversion rate terbesar untuk membeli produk, (perlu dijaga keloyalannya)

Untuk selengkapnya, dapat melihat jupyter notebook disini

Recommendation:

Aktif memonitor transaksi untuk kelompok High Spender, peningkatan service perlu dilakukan
Dilakukan analisis lanjut untuk mid spender agar terjadi peningkatan transaksi dengan memberikan rekomendasi yang tepat
Untuk kelompok Low Spender dan Risk to Churn, dapat dilakukan analisis general, mengingat jumlah visit yang cukup tinggi tapi tidak melakukan transaksi. Mungkin segmentasi promo tidak sesuai.

Potential Impact:

Potential GMV High Spender sebesar IDR 176 Juta, sedangkan untuk kelompok Mid Spender hanya sekitar IDR 66 Juta
Reduksi cost sebesar IDR 50 juta dapat dilakukan apabila dapat melakukan optimasi promo pada kelompok mid spender