Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Muhamad Zamzam Istimaqom
zamzamistimaqom@gmail.com
https://www.linkedin.com/in/zamzamistimaqom

Hi. I am Zamzam Istimaqom,

This report is predict the behaviour of visitor website to predict clicked ads customer classification by using Classification Supervised Machine Learning from EDA to Business Recommendation.

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

Overview

“Sebuah perusahaan di Indonesia ingin mengetahui efektifitas sebuah iklan yang mereka tayangkan, hal ini penting bagi perusahaan agar dapat mengetahui seberapa besar ketercapainnya iklan yang dipasarkan sehingga dapat menarik customers untuk melihat iklan.

Dengan mengolah data historical advertisement serta menemukan insight serta pola yang terjadi, maka dapat membantu perusahaan dalam menentukan target marketing, fokus case ini adalah membuat model machine learning classification yang berfungsi menentukan target customers yang tepat ”

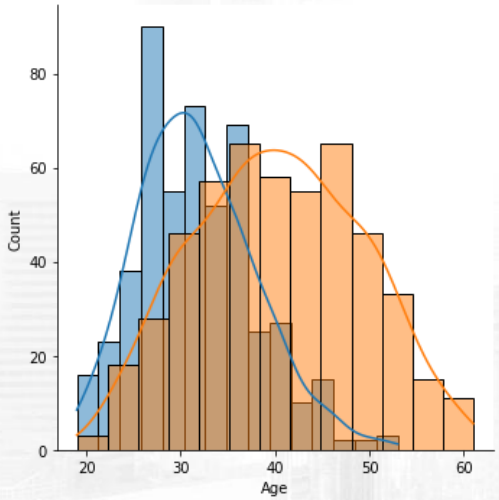
Dataset Overview

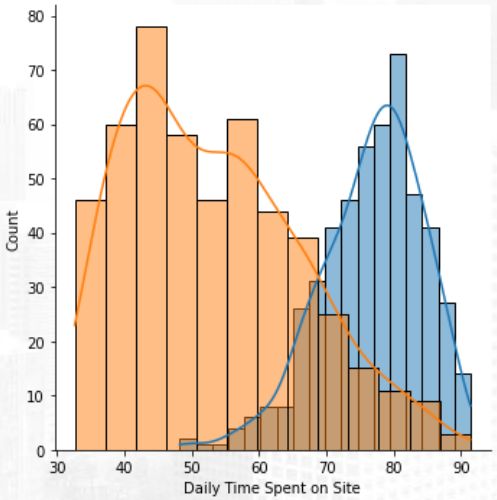
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Unnamed: 0                1000 non-null   int64  
 1   Daily Time Spent on Site  987 non-null    float64
 2   Age                       1000 non-null   int64  
 3   Area Income               987 non-null    float64
 4   Daily Internet Usage      989 non-null    float64
 5   Male                      997 non-null    object 
 6   Timestamp                 1000 non-null   object 
 7   Clicked on Ad             1000 non-null   object 
 8   city                      1000 non-null   object 
 9   province                  1000 non-null   object 
 10  category                  1000 non-null   object 
dtypes: float64(3), int64(2), object(6)
memory usage: 86.1+ KB
```

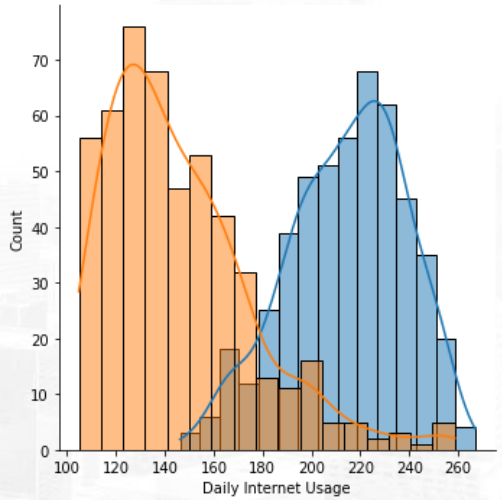
- Data terdiri dari 1000 baris
- Terdapat beberapa missing value pada kolom Daily Time Spent on Site, Area Income, Daily Internet Usage, dan Gender (Male/Female)

Untuk selengkapnya, dapat melihat jupyter notebook disini

Univariate Analysis



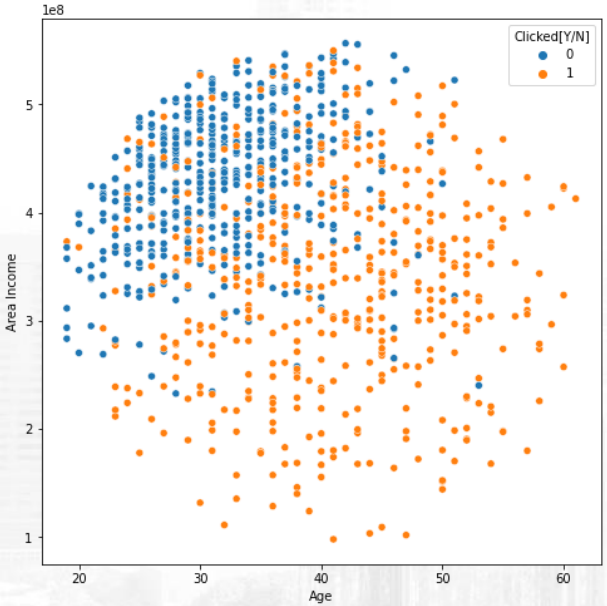


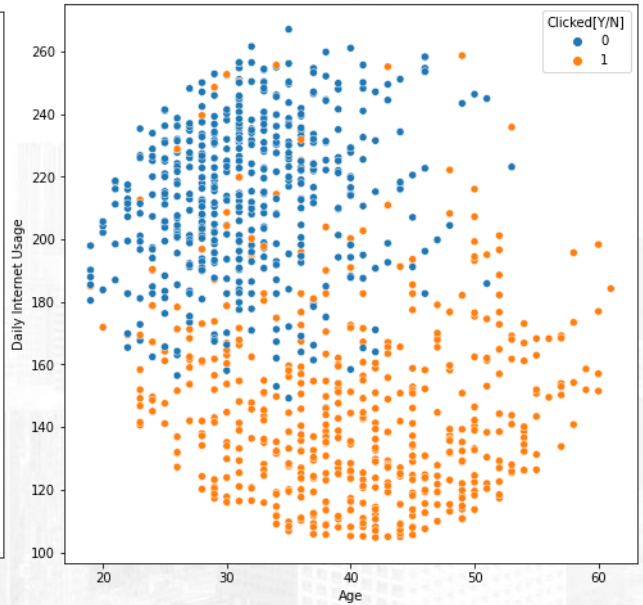


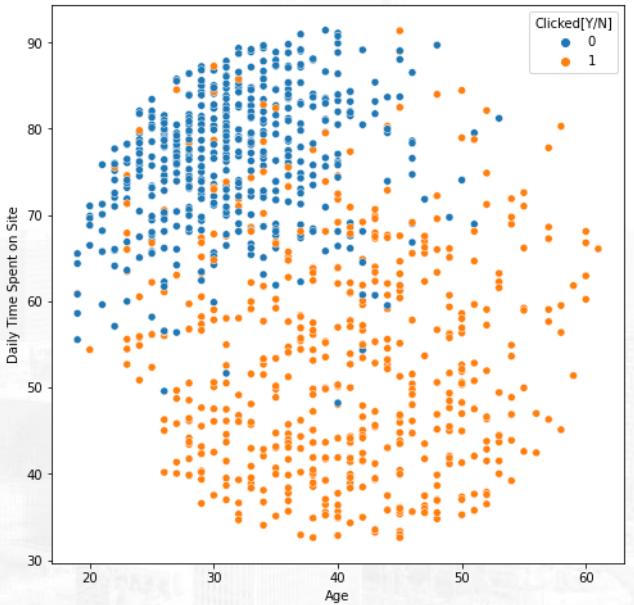
- Rentang usia pengguna internet usia20 - 60 tahun, dan orang dengan usia 40 ke atas cenderung melakukan klik pada iklan yang ditayangkan dibandingkan usia dibawahnya
- Waktu pemakaian penguna internet kurang dari 70 menit per hari cenderung melakukan klik pada iklan. Orang – orang seperti ini dimungkinkan memang mencari informasi yang dibutuhkan sesuai iklan yang dibutuhkan.
- Orang-orang dengan pemakaian internet relatif kecil cenderung efektif melakukan klik iklan dibandingkan yang melakukan penggunaan besar

Untuk selengkapnya, dapat melihat jupyter notebook disini

Bivariate Analysis



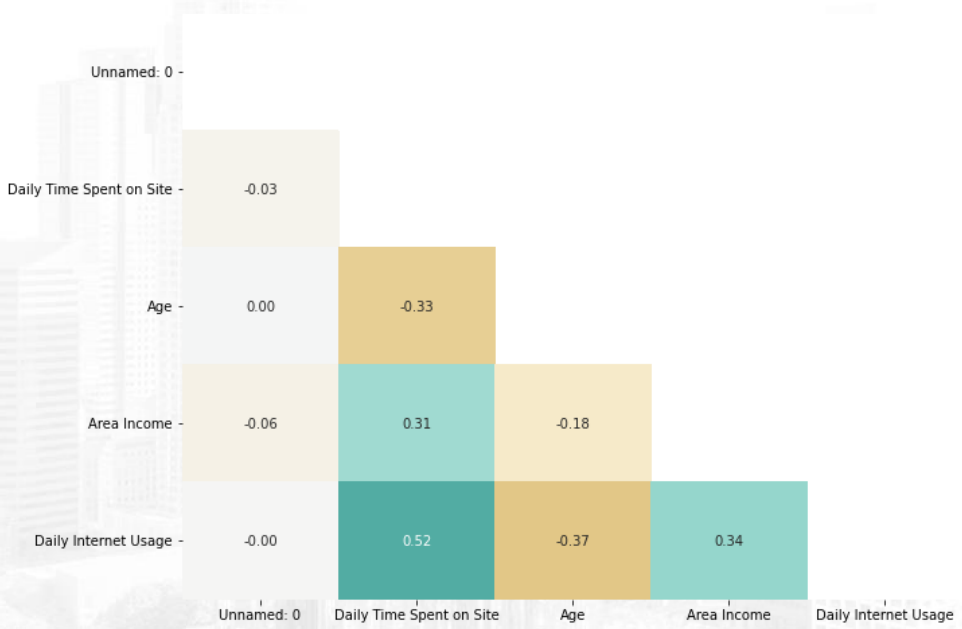


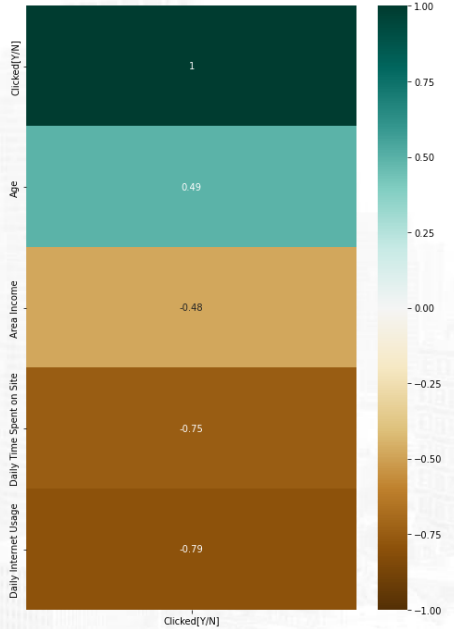


- Persebaran kelompok umur target market secara umum mengarah pada usia 20 – 60 tahun dengan pemakaian internet dan waktu yang sebentar, kemungkinan target yang dituju adalah pekerja produktif yang hanya memiliki waktu sebentar untuk mengakses internet (membuka website) karena mereka bekerja.
- Pengeklik iklan kemungkinan adalah orang dengan penghasilan rendah dengan usia 20 – 40 tahun, sedangkan usia 40 tahun ke atas berpotensi melakukan klik iklan pada usia 40 ke atas.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Heatmap Correlation





- Korelasi umur dengan daily spend time menunjukkan korelasi sedang yaitu 0.33, sehingga semakin tinggi umur semakin tinggi pula penghasilan
- Pemakaian internet harian dan wkatu yang digunakan menunjukkan relasi yang cukup tinggi yaitu 0.52, 2 fitur ini dinilai sangat berkaitan
- Pada Heatmap (bagian kanan), menunjukkan korelasi antar fitur terhadap target (Clicked [Y/N]) dimana Waktu dan pemakaian internet memiliki korelasi kuat

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning & Preprocessing

Missing Value :

Daily Time Spent on Site : 13
Area Income : 13
Daily Internet Usage : 11
Gender : 3

Handle By:

Daily Time Spent on Site : Median ()
Area Income : Median ()
Daily Internet Usage : Median ()
Gender : Mode()

Duplicate Data :

0

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Encoding and Engineering

One Hot Encoding:

0

Label Encoding

Gender (Manual Mapping)
City
Province

Extract Timestamp :

From Timestamp to [‘Year’, ‘Month’, ‘Day’, ‘Hour’]

Untuk selengkapnya, dapat melihat jupyter notebook disini

Train Test Split

Data Split :

Data Training :

80 %

Data Test:

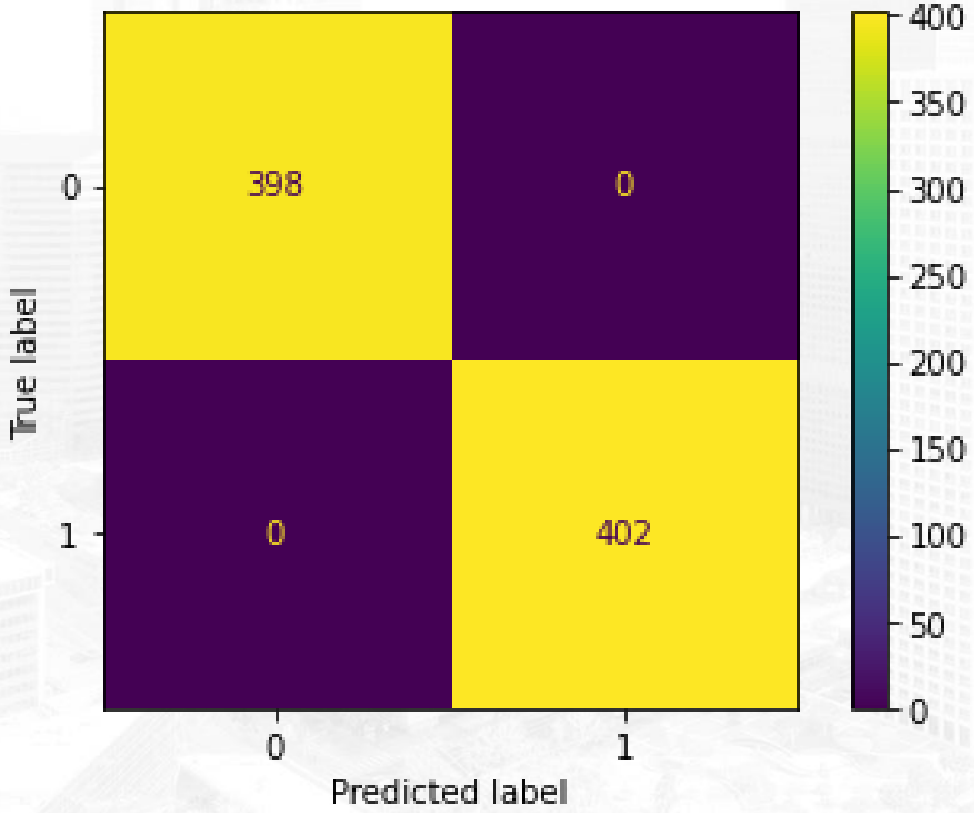
20 %

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling

Model	Accuracy	
	Non Normalisasi	Normalisasi
Logistic Regression	49%	96%
Random Forest	94%	95%
Naïve Baiyes	71%	98%

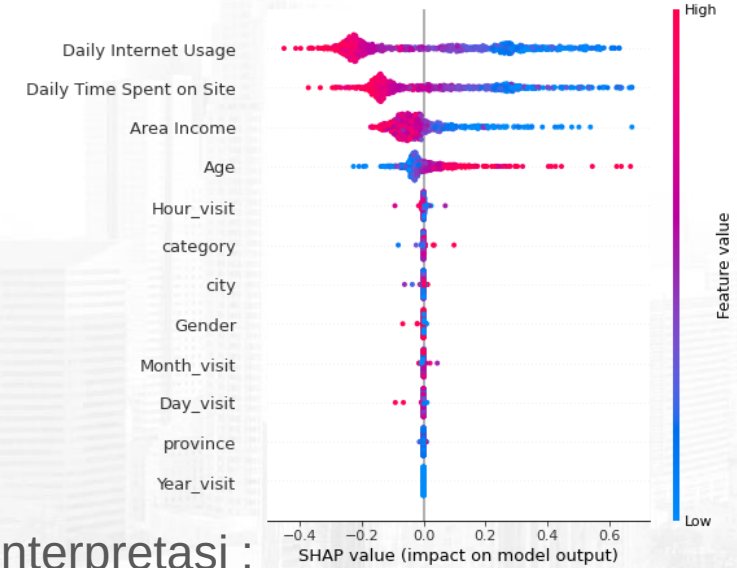
Gause Naïve Bayes Confusion matrix

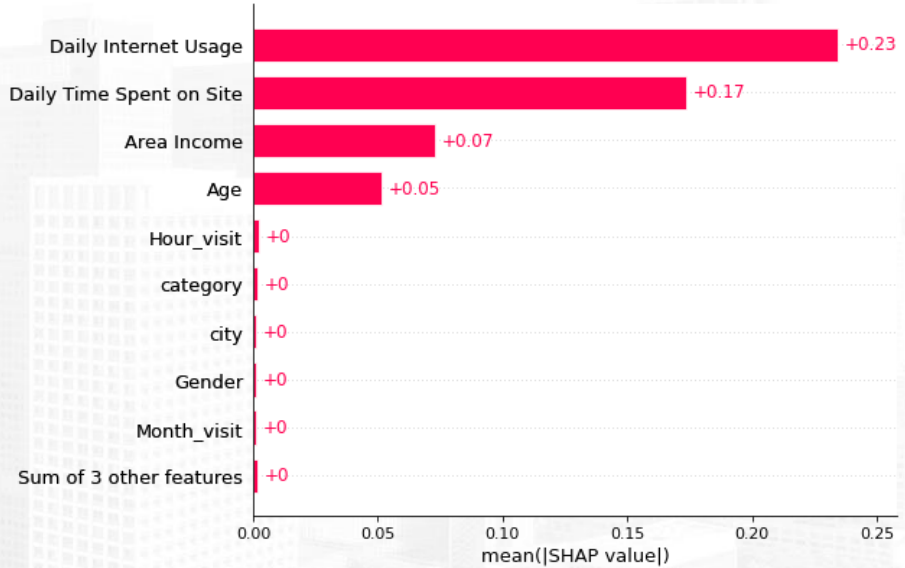


Gnb Classification report

	precision	recall	f1-score	support
0	0.98	0.98	0.98	102
1	0.98	0.98	0.98	98
accuracy			0.98	200
macro avg	0.98	0.98	0.98	200
weighted avg	0.98	0.98	0.98	200

Feature importance Shap Value





Interpretasi :

- Terdapat orang orang yang mengakses internet tidak terlalu lama tetapi memiliki impact pada target dan justru melakukan klik pada iklan
- Orang dengan usia lebih tua cendrung melakukan klik (targeted market)
- Orang dengan penghasilan tinggi, umur tua, dan penggunaan internet sedikit cenderung melakukan klik iklan (targeted market)

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Insight

- Usia yang berpotensi melakukan klik iklan adalah usia 40 tahun ke atas. Hal ini dimungkinkan bahwa produk / layanan sesuai dengan kelompok usia ini. (Rekomendasi : Iklan di set up untuk usia 40 ke atas untuk mengefektifkan budget dan iklan lebih tepat sasaran)
- Target market adalah mereka yang hanya menghabiskan waktu kurang dari 70 menit per hari kemungkinan adalah para pekerja produkif yang hanya sempat mengakses kurang dari durasi waktu tersebut.(Rekomendasi : Membuat campaign iklan dengan keyword yang disesuaikan dengan durasi waktu(promo tertentu untuk melakukan Transaksi dengan Batasan waktu 60 menit agar melakukan konversi)

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Insight

- Machine Learning mampu memprediksi akurasi hingga 98%, sehingga dengan model dapat diketahui behaviour orang orang yang melakukan klik, sehingga dapat berfokus untuk penentuan pengefektifan budget iklan sesuai dengan mayotitas behavior/karakter pengunjung.
- (Rekomendasi : Karena model sudah memiliki akurasi tinggi, behavior orang pada True Negative tidak perlu dijadikan target dan berfokus mengoptimasi iklan berdasarkan karakter utama hasil True Negative yaitu : ‘Orang dengan penghasilan tinggi, umur tua, dan penggunaan internet sedikit cenderung melakukan klik iklan (targeted market)’

Untuk selengkapnya, dapat melihat jupyter notebook disini

Simulasi

Sebelum Menggunakan Machine Learning

- Budget Iklan 3.000.000
- CPM (Cost per million) : 20.000 (Target Impression 150.000)
- Biaya CPC (Cost per click) : 3.000
- Sebelum menggunakan ML dari 1000 orang terdapat 500 Yes dan 500 No (50% Yes)
- Pengunjung 1000 orang dengan klik 500.

Setelah menggunakan machine learning

- CPC diturunkan dan target market di spesifikkan pada kelompok umur
- Setelah menggunakan ML, iklan di optimasi dan difokuskan berdasarkan karakter orang pada True Positive (Memindahkan jumlah pada True Negative ke True Positif) hingga 80%
- Pengunjung 1000 orang dengan jumlah klik 800.

- Jika dihitung dengan budget yang sama yaitu 3.000.000, jumlah klik yang di dapat :
  - Tanpa ML : 10.000 impression (maksimal) + 933 klik
  - Menggunakan ML : 10.000 impression (maksimal) + 1166 klik

Untuk selengkapnya, dapat melihat jupyter notebook disini

Kalkulasi Simulasi

Jika diasumsikan per performance ads terjadi konversi 30% dari perbandingan kedua skema, konversi mendapat Gross Revenue senilai 300 ribu dan profit per produk 50.000

Tanpa ML

- GROSS = 0.3 * 933 * 300.000 = 83.970.000
- PROFIT = 0.3 * 933 * 50.000 = 13.995.000

Menggunakan ML

- GROSS = 0.3 *1166* 300.000 = 104.940.000
- PROFIT = 0.3 *1166* 50.000 = 17.490.000

- Selisih Keuntungan Profit = 3.495.000

(Jika ingin meningkatkan profit, maka yang perlu ditingkatkan budgetingnya)

Kesimpulan

1. Diperlukan set up ulang iklan dengan menargetkan kelompok usia 40 ke atas untuk mengoptimasi iklan
2. Set up ulang iklan pada pekerja dengan penghasilan tinggi dan waktu membuka internet sedikit
3. Profit yang didapatkan tanpa machine learning 13.995.000 sedangkan jika menerapkan hasil menggunakan machine learning naik menjadi 17.490.000, dengan selisih profit sebesar 3.495.000 sehingga didapatkan machine learning mampu memingkatkan penghasilan karena klik yang dihasilkan lebih tertarget