Predict Clicked Ads Customer Classification by using Machine Learning

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

Rakamin Academy

Created by:
Muchammad Malik
muchammad.malik@gmail.com
https://www.linkedin.com/in/muchammad-malik/

“Currently working as business and system development at GESITS, the leading EV brand in Indonesia. Having a bachelor degree from engineering physics , I possessed balanced skill in engineering and management discipline.

As a data science and business analyst enthusiast, I developed skillset in Business Accumen, SQL, Pyhton, Tableau, and machine learning. I have spent 2 years to learn and maintain these skill by taking several bootcamp and online course.

I am a highly-motivated learner to keep me stay relevant, have good analytical thinking, have creative problem solving skill, and able to work in team. During college life, I developed my soft skill by having experienced in leading a commitee, participating in international competition (Model United Nation and business case competition), and also participating in extra-campus organization, such as AIESEC and StudentsCatalyst. ”

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 ”

Customer Type and Behaviour Analysis on Advertisement

- Tulislah proses ***Exploration Data Analysis*** (EDA) yang mencakup ***Statistical analysis*** baik untuk data numerik maupun kategori, Selanjutnya buat visualisasi data untuk ***Univariate*** dan ***Bivariate analysis***, serta ***Multivariate analysis***

- Khusus untuk ***Bivariate analysis***, tunjukan hubungan antara kolom umur, daily internet usage, dan daily time spent on site.

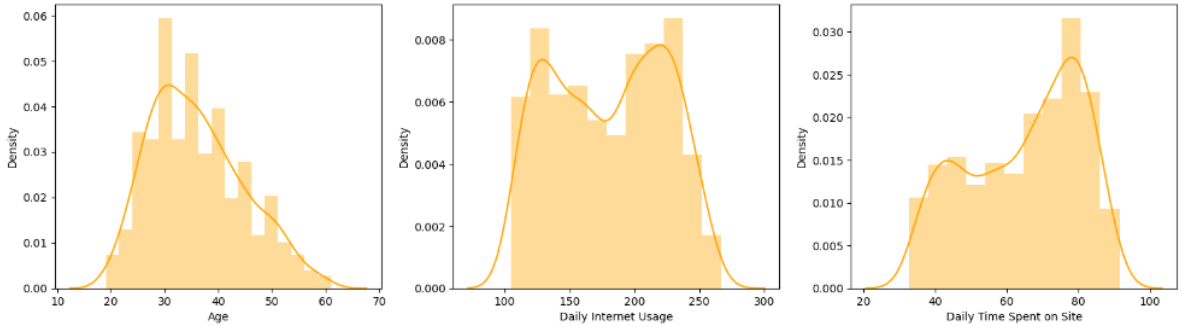
- Tulislah juga **proses korelasi heatmap** untuk mengetahui tingkat korelasi antar kolom

- **Source code** yang sudah kamu buat, dapat ditampilkan dan berikan link untuk mengakses file tersebut. Contohnya seperti di pojok kanan bawah.

Untuk selengkapnya, dapat melihat jupyter notebook disini

UNIVARIATE ANALYSIS

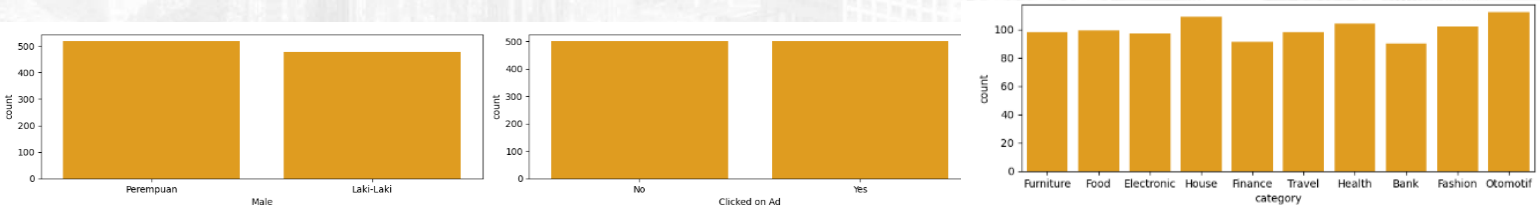
Kolom Numerikal :



Tipe Persebaran Data:

- Feature Umur bersifat positive skew,
- ‘Daily Internet Usage’ bersifat bimodal distribution
- ‘Daily Time Spent on Site’ bersifat negative skew

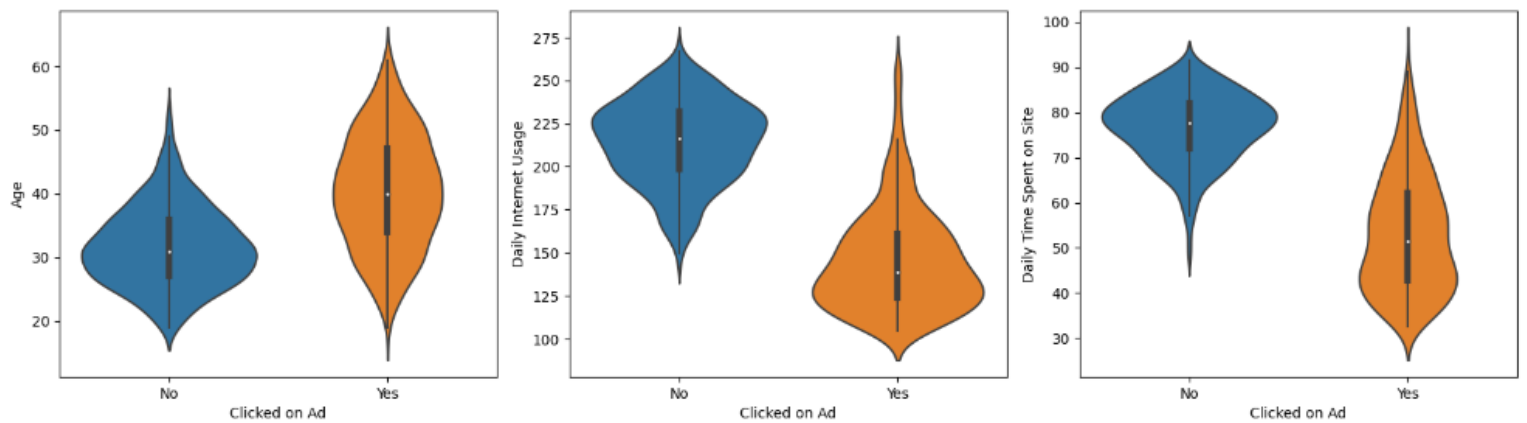
Kolom Kategorikal :



Persebaran data jenis kelamin, kategori produk, dan clicked on ad cenderung berimbang

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/16qQjzyVjFS19EyACopm4FI5bi-7UKbG-?usp=sharing

BIVARIATE ANALYSIS

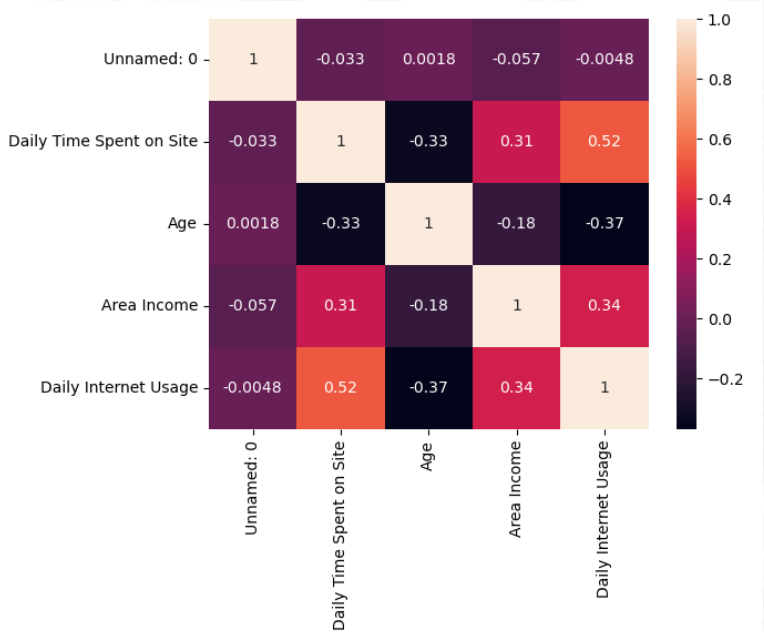


Karakteristik pengunjung website berdasarkan response terhadap iklan:

- Pengunjung yang mengklik iklan memiliki **rata-rata umur lebih tua** dibanding yang mengabaikan iklan
- Pengunjung yang mengklik iklan cenderung **lebih singkat dalam menggunakan internet** dibanding yang mengabaikan iklan
- Pengunjung yang mengklik iklan cenderung **lebih cepat meninggalkan halaman website** dibanding yang mengabaikan iklan

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/16qQjzyVjFS19EyACopm4FI5bi-7UKbG-?usp=sharing

MULTIVARIATE ANALYSIS



Pasangan kolom yang memiliki korelasi yang cukup tinggi diantaranya sebagai berikut:

- Daily Time Spent on Site – Daily Internet Usage
- Age – Daily Internet Usage
- Area Income – Daily Internet Usage

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/16qQjzyVjFS19EyACopm4FI5bi-7UKbG-?usp=sharing

Data Cleaning & Preprocessing

- Pada tahap ***cleaning data***, tunjukan ***null*** atau ***missing value*** serta ***duplicated value*** pada dataset, serta cara penyelesaiannya.

- Tulislah pula proses **extract datetime data** sebelum dilakukan model machine learning.

- Tunjukan **Split Data** sebelum melakukan model machine learning

- Tulislah proses **feature encoding** pada tahap ini (gunakan get_dummy)

- **Source code** yang sudah kamu buat, dapat ditampilkan dan berikan link untuk mengakses file tersebut. Contohnya seperti di pojok kanan bawah.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Cleaning & Preprocessing

CLEANING DATA

Duplicated data: Tidak ada

Missing data:

Daily Time Spent on Site	13
Age	0
Area Income	13
Daily Internet Usage	11
Male	3
Timestamp	0
Clicked on Ad	0
city	0
province	0
category	0

Cara handle missing data:

- Pada kolom 'Male': missing data dihilangkan karena jumlahnya terlalu sedikit
- Pada kolom 'Daily Time Spent on Site : diganti dengan nilai mean
- Pada kolom 'Area Income' : diganti dengan nilai mean
- Pada kolom 'Daily Internet Usage' : diganti dengan nilai mean

FEATURE ENCODING

Kolom Male
Label encoding dengan perubahan value
Laki-laki → 1
Perempuan → 0

Kolom Clicked on Ad
Label encoding dengan perubahan value
Yes → 1
No → 0

Kolom City — One-Hot Encoding

Kolom Province — One-Hot Encoding

Kolom Category — One-Hot Encoding

city_Batam	city_Bekasi
0	0
0	0
0	0
1	0
0	0

prov_Bali	prov_Banten	prov_Daerah Khusus Ibukota Jakarta
0	0	1
1	0	0
0	0	0
0	0	0
0	0	0

cat_Electronic	cat_Fashion
0	0
0	0
1	0
0	0
0	0

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/16qQjzyVjFS19EyACopm4FI5bi-7UKbG-?usp=sharing

EXTRACT DATETIME DATA

```
df['Timestamp'] = pd.to_datetime(df['Timestamp'])

df['Year_clicked'] = df['Timestamp'].apply(lambda x: x.year)
df['Month_clicked'] = df['Timestamp'].apply(lambda x: x.month)
df['Week_clicked'] = df['Timestamp'].apply(lambda x: x.week)
df['Day_clicked'] = df['Timestamp'].apply(lambda x: x.day)

df.drop('Timestamp', axis=1, inplace=True)
```

Perubahan tipe data kolom Timestamp dari tipe data string menjadi tipe data Datetime

Ekstraksi kolom Timestamp menjadi tahun, bulan, minggu, dan hari

Drop kolom Timestamp karena sudah tidak dibutuhkan lagi

SPLIT DATA

```
#kolom fitur
X = df.drop('Clicked on Ad', axis=1)

#kolom target
y = df['Clicked on Ad']
```

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/16qQjzyVjFS19EyACopm4FI5bi-7UKbG-?usp=sharing

Data Modeling

- Tulislah proses model machine learning terdiri dari
    a. **hasil *experiment 1*** (sebelum normalisasi/standardisasi),
    b. **hasil *experiment 2*** (setelah normalisasi/standardisasi).
    c. hasil tabel **confusion matrix** dari model tersebut.
    d. Daftar ***Feature Important***.

- Tulislah hasil interpretasi dari model tersebut

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Modeling

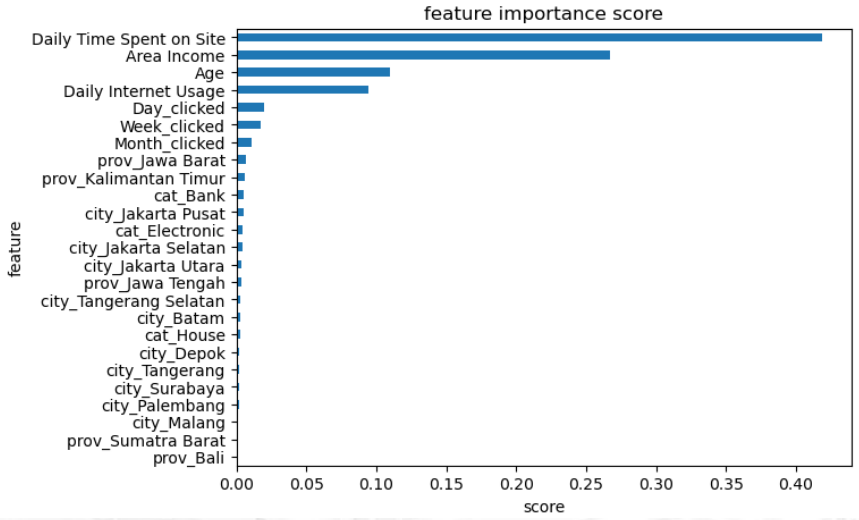
	Hasil Experiment 1 – Tanpa Normalisasi	Hasil Experiment 2 – Dengan Normalisasi
Confusion Matrix	[[88 23] [9 80]]	[[90 21] [6 83]]
Metriks Evaluasi	Accuracy (Test Set): 0.84 Accuracy (Train Set): 0.94 Precision (Test Set): 0.78 Precision (Train Set: 0.94 Recall (Test Set): 0.90 Recall (Train Set): 0.93 F1-Score (Test Set): 0.83 F1-Score (Train Set): 0.94 roc_auc (test-proba): 0.87 roc_auc (train-proba): 0.99 recall (crossval train): 0.9018558897243107 recall (crossval test): 0.7615353535353535	Accuracy (Test Set): 0.86 Accuracy (Train Set): 0.94 Precision (Test Set): 0.80 Precision (Train Set: 0.94 Recall (Test Set): 0.93 Recall (Train Set): 0.94 F1-Score (Test Set): 0.86 F1-Score (Train Set): 0.94 roc_auc (test-proba): 0.88 roc_auc (train-proba): 0.99 recall (crossval train): 0.8612443609022555 recall (crossval test): 0.7875757575757575

Interpretasi:
Setelah dilakukan normalisasi, overfitting dapat dikurangi. Hal ini dilihat dari selisih metrics pada data train dan data test yang semakin sedikit

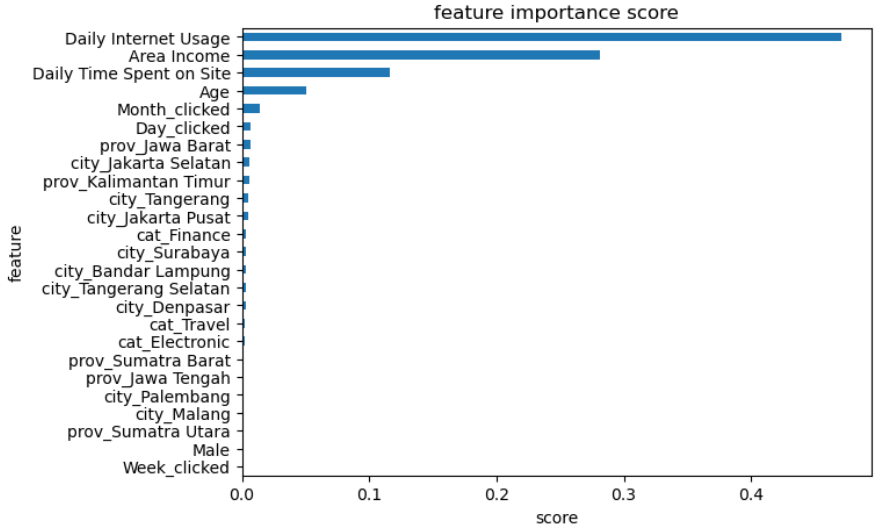
Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/16qQjzyVjFS19EyACopm4FI5bi-7UKbG-?usp=sharing

Feature Importance

Hasil Experiment 1 – Tanpa Normalisasi



Hasil Experiment 2 – Dengan Normalisasi



Interpretasi:

- Empat feature terpenting pada eksperimen tanpa normalisasi dan dengan normalisasi adalah sama.
- Perbedaannya terletak pada urutan fitur ketiga dan fitur keempat

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/16qQjzyVjFS19EyACopm4FI5bi-7UKbG-?usp=sharing

Business Recommendation & Simulation

- Tampilkan ***Feature Important*** dari hasil model machine learning

- Tulislah **rekomendasi bisnis** berdasarkan EDA dan Feature Important

- Tulislah sebuah simulasi perusahaan dalam marketing yang menunjukan ***cost, revenue, dan profit*** **sebelum dan setelah menggunakan model machine learning**. Tunjukan perbedaan dari kedua simulasi tersebut.

- Tulislah pula **simpulan** yang didapat dari proses tersebut

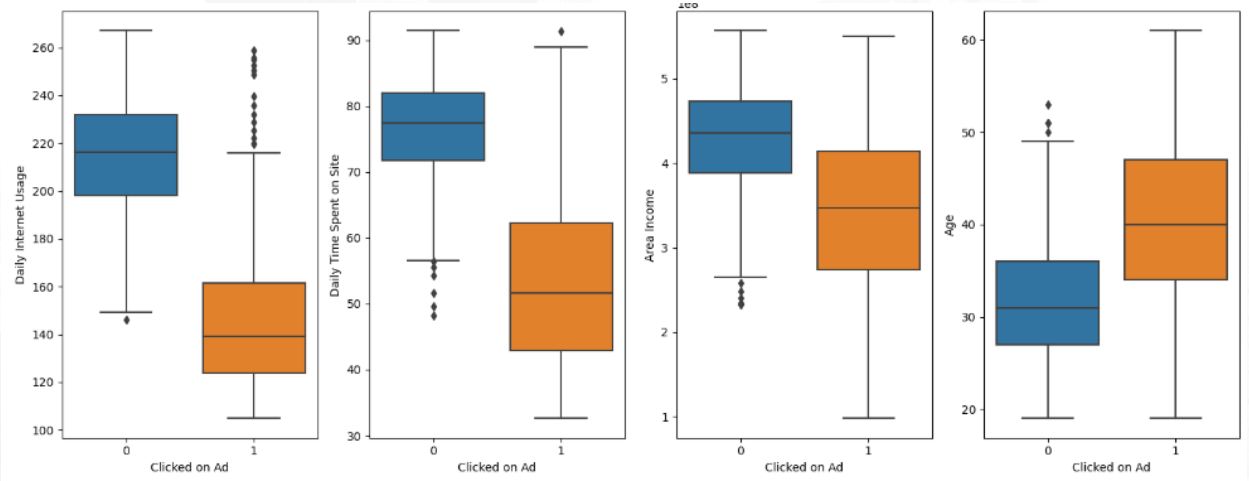
Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation

Visualisasi EDA dari 4 fitur terpenting

Top 4 Features

Daily Internet Usage	0.470948
Area Income	0.281165
Daily Time Spent on Site	0.116519
Age	0.050436



Rekomendasi Bisnis

Iklan sebaiknya ditargetkan kepada customer yang:

- Daily Internet usage di antara 120-165 menit per hari. Semakin kecil daily internet usage, semakin besar kemungkinan untuk klik ad
- Daily Time Spent on Site di antara 40-65 menit per hari. Semakin kecil daily time spent on site, semakin besar kemungkinan untuk klik ad
- Area Income diantara 2,5-4,5. Semakin kecil Income, semakin besar kemungkinan untuk klik ad
- Usia diantara 35-50 tahun. Semakin tua usia, semakin besar kemungkinan untuk klik ad

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/16qQjzyVjFS19EyACopm4FI5bi-7UKbG-?usp=sharing

Simulasi Bisnis

No	Metriks Pembanding	Tanpa Modelling	Dengan Modelling
0	Jumlah Campaign	997	104
1	Jumlah Klik	499	83
2	Revenue	99.800.000	16.600.000
3	Cost	99.700.000	10.400.000
4	Profit	100.000	6.200.000

Untuk selengkapnya, dapat melihat jupyter notebook disini:
https://colab.research.google.com/drive/16qQjzyVjFS19EyACopm4FI5bi-7UKbG-?usp=sharing