Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Limatan Luviar
Limatan.junior@gmail.com
Limatan Luviar

Graduated from Universitas Sriwijaya majoring in Informatics Management in 2018 and interned at a government institution. Enrolling in and completing the Data Science Bootcamp at Rakamin Academy in 2024 was driven by my interest in data analysis. Currently, I am actively seeking job opportunities and have completed various data-driven projects at Rakamin. My focus is on Data Analyst roles, where I can utilize my analytical skills and business understanding to make a significant impact

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 ”

Exploratory Data Analysis

	Feature	Data Type	Null	Null (%)	Unique	Unique Sample
0	Unnamed: 0	int64	0	0.0	1000	[0, 1, 2, 3, 4]
1	Daily Time Spent on Site	float64	13	1.3	890	[68.95, 80.23, 69.47, 74.15, 68.37]
2	Age	int64	0	0.0	43	[35, 31, 26, 29, 23]
3	Area Income	float64	13	1.3	987	[432837300.0, 479092950.00000006, 418501580.0,...
4	Daily Internet Usage	float64	11	1.1	955	[256.09, 193.77, 236.5, 245.89, 225.58]
5	Male	object	3	0.3	2	[Perempuan, Laki-Laki, nan]
6	Timestamp	object	0	0.0	997	[3/27/2016 0:53, 4/4/2016 1:39, 3/13/2016 20:3...
7	Clicked on Ad	object	0	0.0	2	[No, Yes]
8	city	object	0	0.0	30	[Jakarta Timur, Denpasar, Surabaya, Batam, Medan]
9	province	object	0	0.0	16	[Daerah Khusus Ibukota Jakarta, Bali, Jawa Tim...
10	category	object	0	0.0	10	[Furniture, Food, Electronic, House, Finance]

Berdasarkan hasil tersebut, terdapat beberapa insight yang dapat diperoleh:

- Null Values:

Daily Time Spent on Site, Area Income, Male dan Daily Internet Usage memiliki nilai null. Dalam hal ini, persentase nilai null cukup kecil (kurang dari 5%), sehingga tidak akan terlalu mempengaruhi hasil analisis secara signifikan.

- Data Types

Feature Male memiliki tipe data object Sebaiknya diubah menjadi tipe data yang sesuai, seperti boolean (0 atau 1) atau kategorikal. Feature Timestamp lebih baik menggunakan type data datetime

- Unique Values

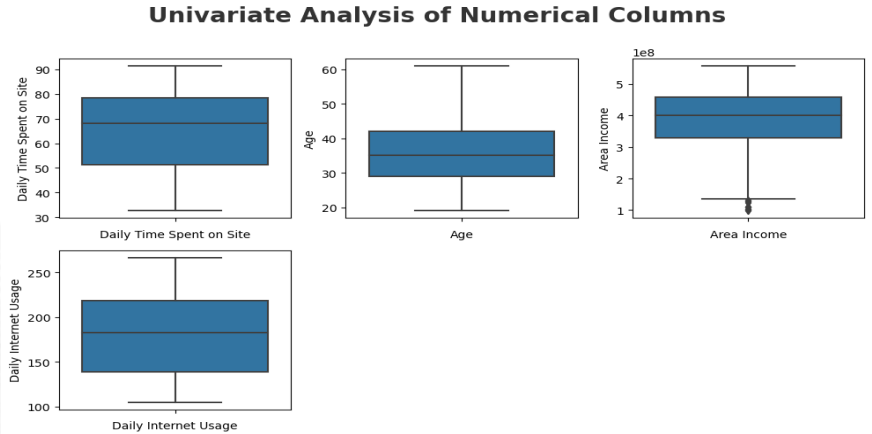
Feature Unnamed: 0 memiliki 1000 nilai unique.** Hal ini menunjukkan bahwa kolom ini mungkin merupakan indeks atau nomor baris yang tidak memberikan informasi yang berguna untuk analisis. Sebaiknya dihapus dari dataset.

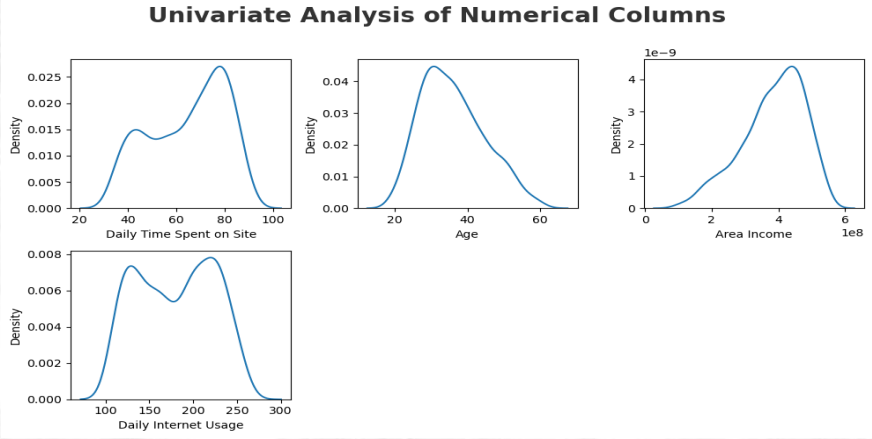
- Target Feature

Feature Clicked on Ad mungkin menjadi target dalam analisis.** menjadi target feature.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Exploratory Data Analysis

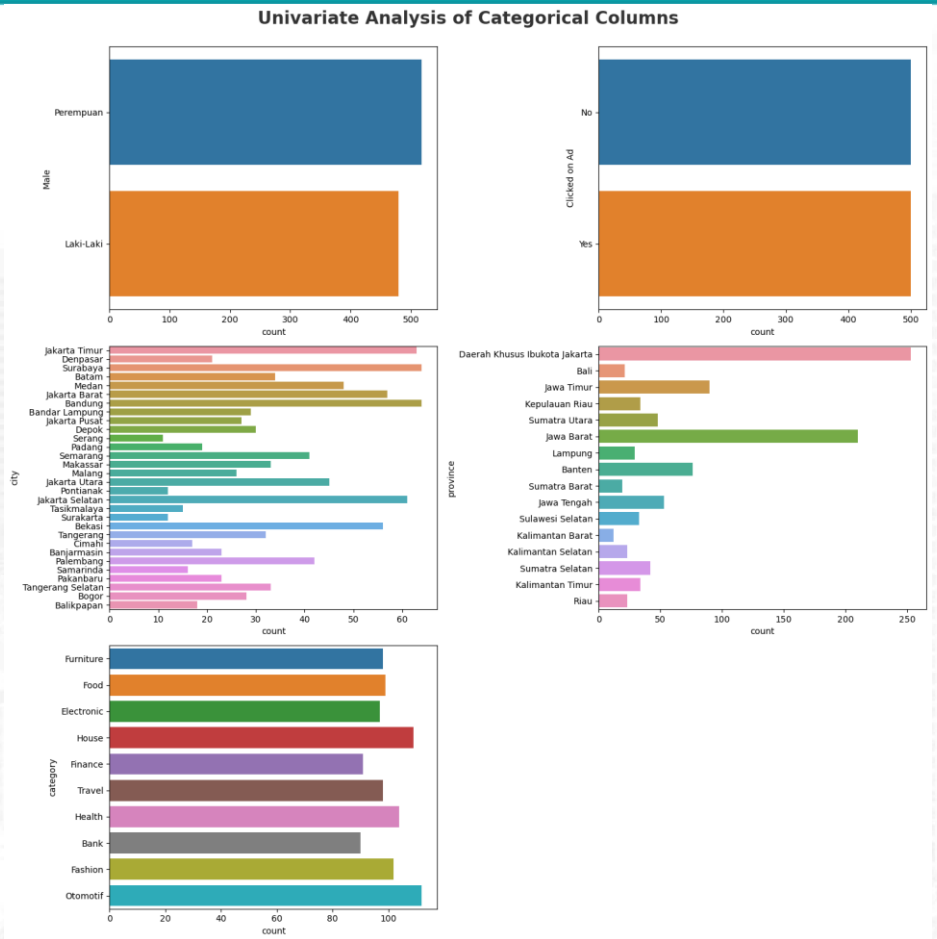




- Daily Time Spent on Site: Rata-rata waktu yang dihabiskan harian di situs adalah sekitar 64.93 menit, dengan standar deviasi sekitar 15.84 menit. Waktu minimum yang dihabiskan adalah 32.60 menit dan maksimumnya adalah 91.43 menit. Dan distribusinya terlihat bimodal
- Age: Rata-rata usia responden adalah sekitar 36 tahun, dengan standar deviasi sekitar 8.79 tahun. Usia minimum adalah 19 tahun dan maksimumnya adalah 61 tahun. dan distribusinya terlihat hampir normal
- Area Income: Rata-rata pendapatan area responden adalah sekitar 384,864,700, dengan standar deviasi sekitar 94,079,990. Pendapatan area minimum adalah 97,975,500 dan maksimumnya adalah 556,393,600. dan distribusinya terlihat skewed ke kanan (positive)
- Daily Internet Usage: Rata-rata penggunaan internet harian adalah sekitar 179.86 MB, dengan standar deviasi sekitar 43.87 MB. Penggunaan internet harian minimum adalah 104.78 MB dan maksimumnya adalah 267.01 MB. Dan distribusinya terlihat bimodal

Untuk selengkapnya, dapat melihat jupyter notebook disini

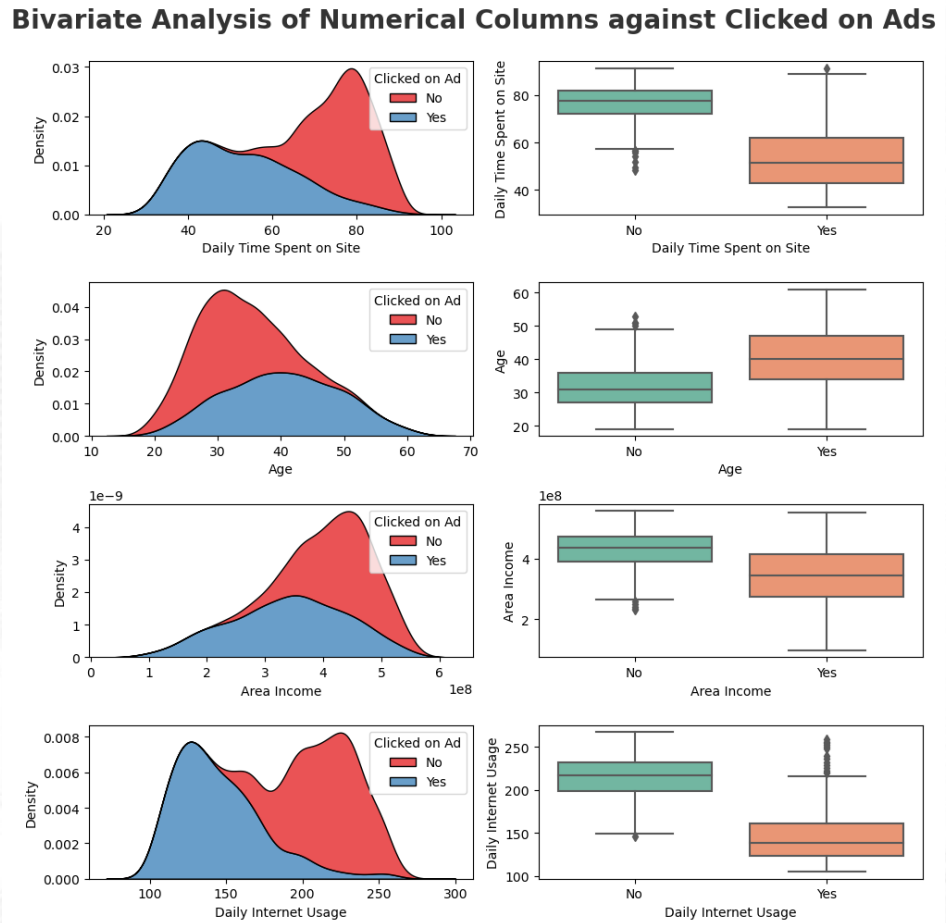
Exploratory Data Analysis



- Male (Jenis Kelamin): Terdapat 997 entri dalam kolom ini, dengan 2 nilai unik yaitu "Perempuan" dan "Laki-laki". Nilai yang paling sering muncul (mode) adalah "Perempuan" dengan frekuensi 518.
- Clicked on Ad (Klik pada Iklan): Terdapat 1000 entri dalam kolom ini, dengan 2 nilai unik yaitu "Yes" dan "No". Nilai yang paling sering muncul adalah "No" dengan frekuensi 500.
- City (Kota): Terdapat 1000 entri dalam kolom ini, dengan 30 nilai unik yang mewakili nama-nama kota. Kota "Surabaya" adalah yang paling sering muncul dengan frekuensi 64.
- Province (Provinsi): Terdapat 1000 entri dalam kolom ini, dengan 16 nilai unik yang mewakili nama-nama provinsi. Provinsi "Daerah Khusus Ibukota Jakarta" adalah yang paling sering muncul dengan frekuensi 253.
- Category (Kategori): Terdapat 1000 entri dalam kolom ini, dengan 10 nilai unik yang mewakili kategori-kategori tertentu. Kategori "Otomotif" adalah yang paling sering muncul dengan frekuensi 112.

Untuk selengkapnya, dapat melihat jupyter notebook disini

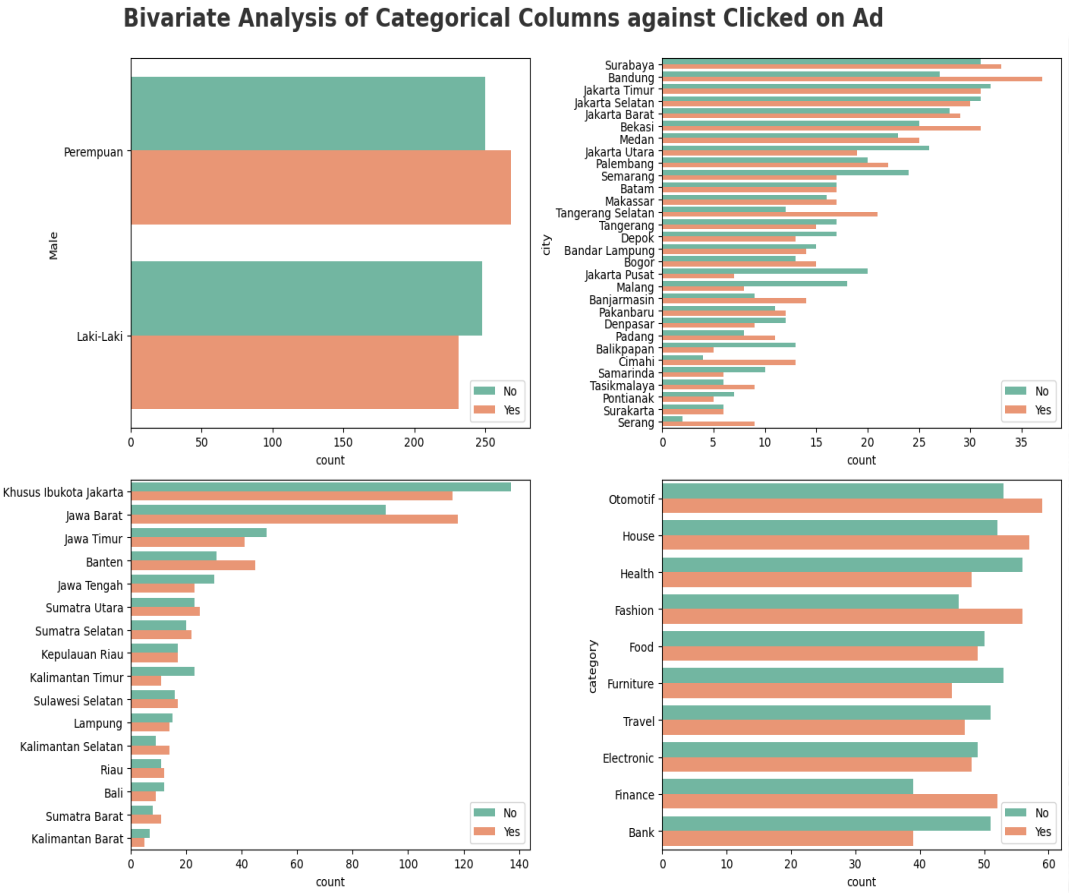
Exploratory Data Analysis



- Daily Time Spent on Site: Pengguna yang tidak Clicked on Ad memiliki rata-rata waktu yang lebih lama di situs (76.79 menit) dibandingkan dengan pengguna yang Clicked on Ad (53.14 menit). Median waktu yang dihabiskan juga menunjukkan pola yang sama.
- Age: Rata-rata usia pengguna yang tidak Clicked on Ad (31.68 tahun) lebih rendah dibandingkan dengan pengguna yang Clicked on Ad (40.33 tahun). Median usia juga menunjukkan pola yang sama.
- Area Income: Rata-rata dan median Area Income pengguna yang tidak Clicked on Ad lebih tinggi daripada pengguna yang Clicked on Ad. Ini menunjukkan bahwa pengguna dengan pendapatan area yang lebih tinggi cenderung untuk tidak Clicked on Ad.
- Daily Internet Usage: Pengguna yang tidak Clicked on Ad memiliki rata-rata penggunaan internet harian yang lebih tinggi (214.60 MB) dibandingkan dengan pengguna yang Clicked on Ad (145.34 MB). Median penggunaan internet harian juga menunjukkan pola yang sama.

Untuk selengkapnya, dapat melihat jupyter notebook disini

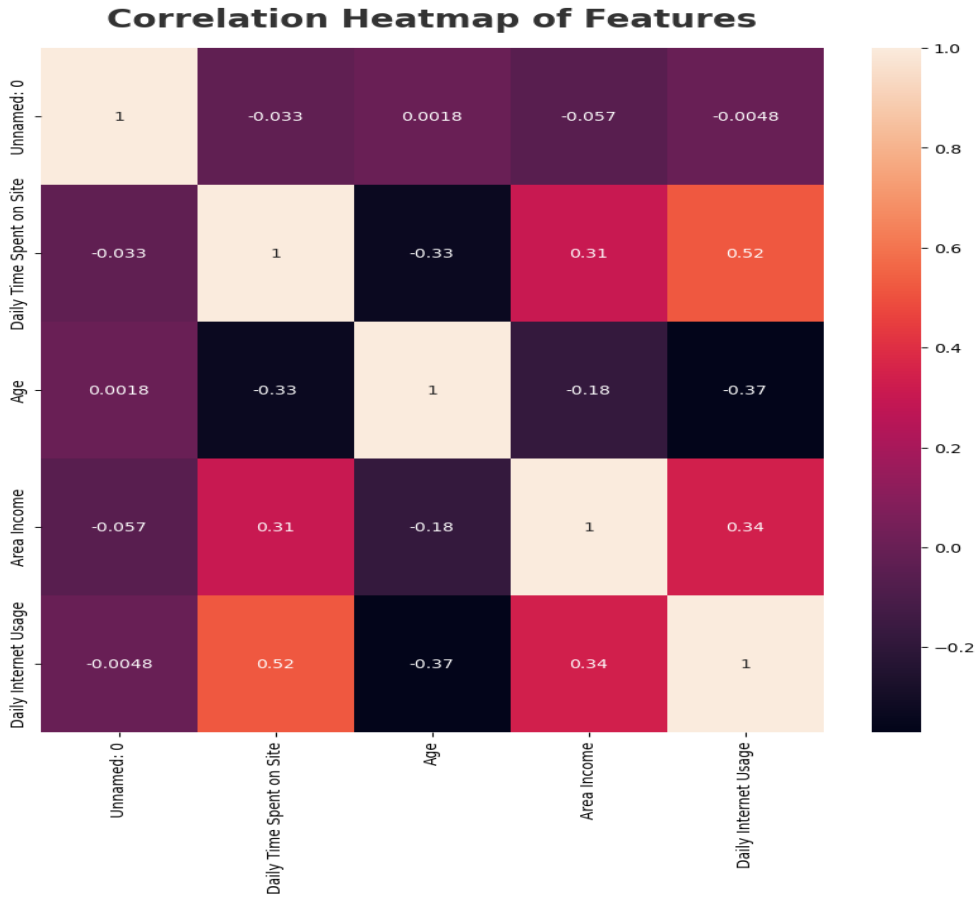
Exploratory Data Analysis



- Lebih banyak pengguna perempuan yang Clicked on Ads dibandingkan pengguna laki-laki pada fitur Male.
- Kota Bandung memiliki jumlah pengguna yang paling banyak Clicked on Ads.
- Pengguna di Provinsi Jawa Barat cenderung untuk Clicked on Ads yang diberikan, sementara pengguna di DKI Jakarta lebih banyak yang menghiraukan iklan.
- Kategori Otomotif cenderung memiliki lebih banyak klik, diikuti oleh Fashion dan House. Sementara kategori Health memiliki jumlah tertinggi untuk tidak Clicked on Ads.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Exploratory Data Analysis



- Daily Time Spent user menghabiskan waktu di situs, semakin tinggi Daily Internet Usage mereka.
- Orang dengan area income yang lebih tinggi cenderung menghabiskan lebih banyak waktu di situs dan memiliki Daily Internet Usage yang lebih tinggi.
- Semakin tua seseorang, kemungkinan mereka Daily Time Spent di situs web cenderung lebih sedikit dan Daily Internet Usage mereka cenderung sedikit lebih rendah.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

Handle Missing Value

	Jumlah Missing Value	Persentase (%)
Daily Time Spent on Site	13	1.3
Area Income	13	1.3
Daily Internet Usage	11	1.1
Male	3	0.3

- 13 nilai hilang pada kolom "Daily Time Spent on Site" (sekitar 1.3% dari total data), diimputasi dengan nilai median.
- 13 nilai hilang pada kolom "Area Income" (sekitar 1.3% dari total data), diimputasi dengan nilai median.
- 11 nilai hilang pada kolom "Daily Internet Usage" (sekitar 1.1% dari total data), diimputasi dengan nilai median.
- 3 nilai hilang pada kolom "Male" (sekitar 0.3% dari total data), diimputasi dengan modus.

Handle Duplicated value

```
df_prep.duplicated().sum()
✓ 0.0s
0
```

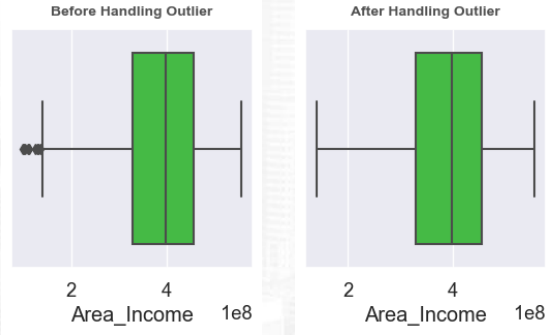
Tidak ada data yang duplikat jadi kita akan melewati Langkah ini

Feature Engineering

```
# Convert 'Timestamp' column to datetime format
df_prep['Date'] = pd.to_datetime(df_prep['Date'])
df_prep['Year'] = df_prep.Date.dt.year
df_prep['Month'] = df_prep.Date.dt.month
df_prep['Week'] =df_prep.Date.dt.dayofweek
df_prep['Day'] = df_prep.Date.dt.day
```

Membuat feature baru dari Feature Date yaitu, Year, Month, Week, Day

Handle Outlier



Terdapat Outlier pada feature Area_Income, Handing Outlier menggunakan IQR

Untuk selengkapnya, dapat melihat jupyter notebook disini

Data Preprocessing

Feature Encoding

```
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import LabelEncoder

# Inisialisasi LabelEncoder
label_encoder = LabelEncoder()
# Melakukan encoding pada fitur 'gender' dengan LabelEncoder
df_prep['Gender_encoded'] = label_encoder.fit_transform(df_prep['Gender'])

# Melakukan encoding pada fitur 'Clicked on Ad' dengan LabelEncoder
df_prep['clickedads_encoded'] = label_encoder.fit_transform(df_prep['Clicked on Ad'])

# # Inisialisasi OneHotEncoder
onehot_encoder = OneHotEncoder()
# handle dengan one hot encoding
for cat in ['city', 'province', 'category']:
  onehots = pd.get_dummies(df_prep[cat], prefix=cat)
  df_prep = df_prep.join(onehots)
```

- Encoding menggunakan One Hot Encoding pada feature city, province, dan category
- Label Encoding pada feature Gender & Clicked on Ad

Feature Selection

Feature Selection dengan menghapus kolom-kolom yang tidak diperlukan untuk analisis selanjutnya. Kolom-kolom yang dihapus adalah 'Unnamed: 0', 'Date', 'Clicked on Ad', 'city', 'province', 'category', dan 'Gender'. Hal ini dilakukan untuk menyederhanakan dataset dan memfokuskan analisis pada fitur-fitur yang lebih relevan.

Split Train & Test Data

```
from sklearn.model_selection import train_test_split
X = df_model.drop(labels=['clickedads_encoded'],axis=1)
y = df_model[['clickedads_encoded']]

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size = 0.3,stratify=y,random_state = 42)
print('Train:',X_train.shape)
print('Test:',X_test.shape)
```

```
✓ 0.0s

Train: (700, 65)
Test: (300, 65)
```

- Membagi feature menjadi target dan feature
- Melakukan pembagian data train dan data test dengan pembagian 70:30

Untuk selengkapnya, dapat melihat jupyter notebook disini

MODELLING TANPA NORMALIZATION

	model_name	model	accuracy	recall	precision	f1_score	duration
0	K-Nearest Neighbor	KNeighborsClassifier()	0.673333	0.640000	0.685714	0.662069	0.009973
1	XgBoost	XGBClassifier(base_score=None, booster=None, c...	0.963333	0.966667	0.960265	0.963455	0.345587
2	Random Forest	(DecisionTreeClassifier(max_features='sqrt', r...	0.963333	0.973333	0.954248	0.963696	0.631311
3	Gradient Boosting	([DecisionTreeRegressor(criterion='friedman_ms...	0.960000	0.960000	0.960000	0.960000	1.185767
4	LightGBM	LGBMClassifier(force_col_wise=True)	0.973333	0.966667	0.979730	0.973154	0.175537

Model LightGBM memiliki performa yang paling baik dibandingkan dengan model lainnya untuk dataset dan metrik evaluasi yang digunakan. Hal ini dapat dilihat dari nilai accuracy, recall, precision, dan f1_score yang tertinggi dibandingkan dengan model lain

Untuk selengkapnya, dapat melihat jupyter notebook disini

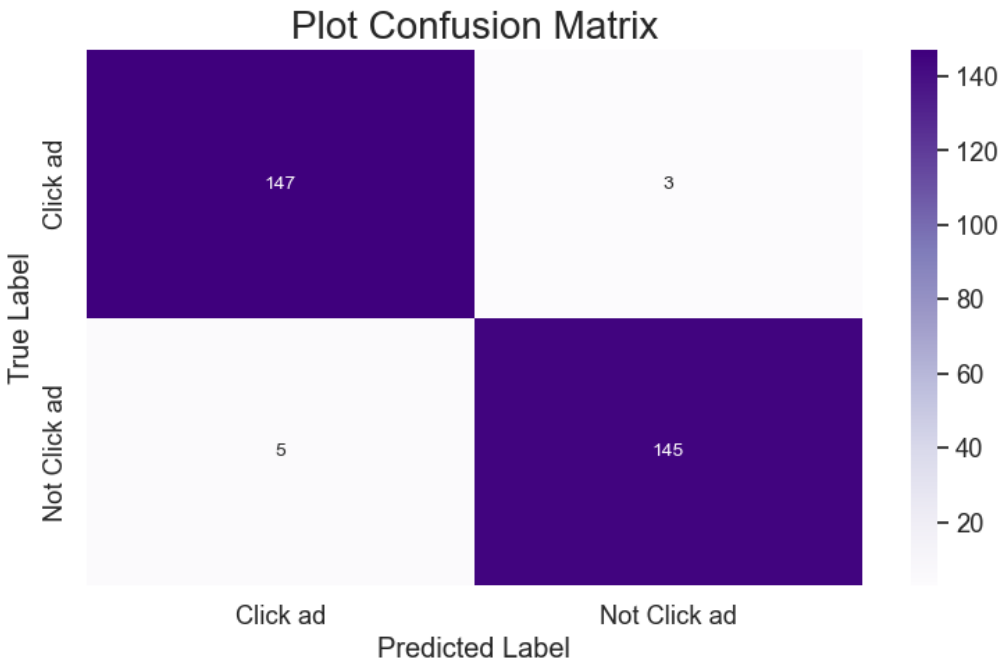
MODELLING DENGAN NORMALIZATION

	model_name	model	accuracy	recall	precision	f1_score	duration
0	K-Nearest Neighbor	KNeighborsClassifier()	0.723333	0.686667	0.741007	0.712803	0.004988
1	XgBoost	XGBClassifier(base_score=None, booster=None, c...	0.963333	0.966667	0.960265	0.963455	0.148604
2	Random Forest	(DecisionTreeClassifier(max_features='sqrt', r...	0.966667	0.966667	0.966667	0.966667	0.524598
3	Gradient Boosting	([DecisionTreeRegressor(criterion='friedman_ms...	0.960000	0.960000	0.960000	0.960000	1.036226
4	LightGBM	LGBMClassifier(force_col_wise=True)	0.970000	0.966667	0.973154	0.969900	0.168556

- Model K-Nearest Neighbor mengalami peningkatan performa
- Model XGBoost, Random Forest, dan Gradient Boosting memiliki performa yang relatif sama dengan sebelum normalisasi, tetapi LightGBM mengalami sedikit penurunan performa, namun LightGBM masih menjadi model yang terbaik dengan accuracy, recall, precision dan f1 score tertinggi.

Untuk selengkapnya, dapat melihat jupyter notebook disini

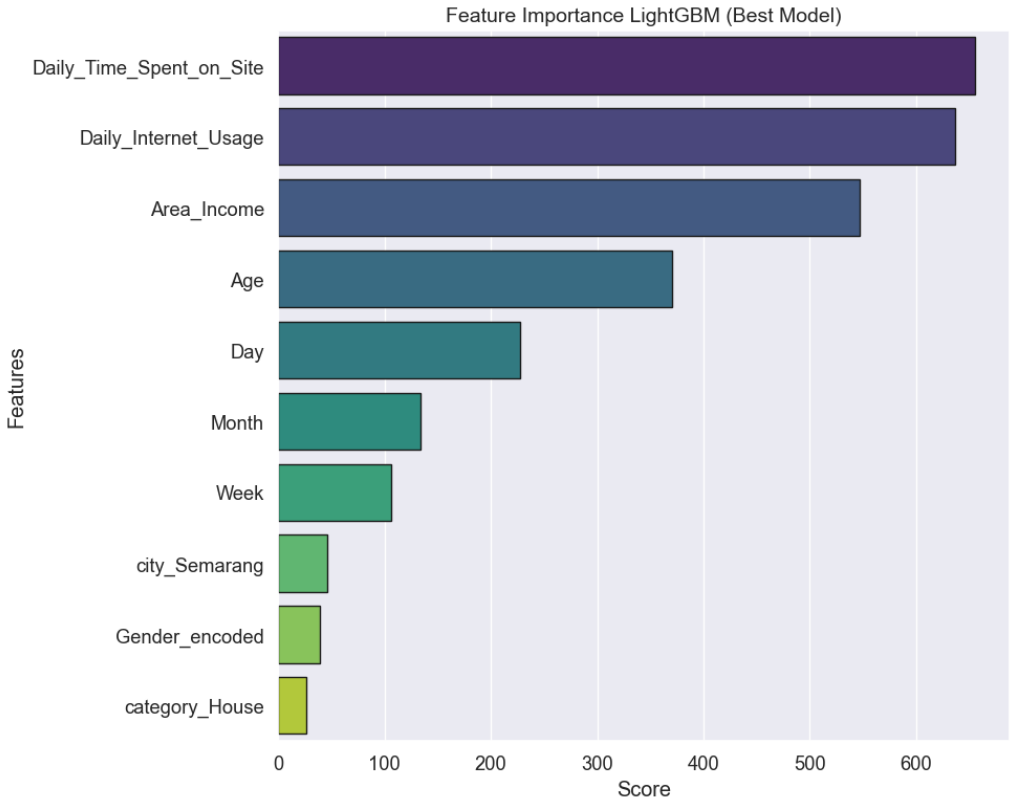
Confusion Matrix



- Confusion matrix memberikan gambaran performa model dalam membedakan kelas "Click ad" dan "Not Click ad", mengidentifikasi kesalahan, dan menentukan strategi perbaikan. Dalam konteks ini.
- mengurangi False Negative dan False Positive penting untuk meningkatkan akurasi model.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Feature Importance



1. **Daily_Time_Spent_on_Site**: Fitur ini memiliki nilai feature importance tertinggi, menunjukkan bahwa waktu harian yang dihabiskan pengguna di situs web memiliki pengaruh paling besar terhadap prediksi.
2. **Daily_Internet_Usage**: memiliki pengaruh yang signifikan terhadap prediksi.
3. **Area_Income**: Pendapatan area tempat pengguna berada juga berkontribusi besar terhadap prediksi.
4. **Age**: Usia pengguna mempengaruhi prediksi dalam jumlah yang cukup signifikan.
5. **Day**: Hari dalam sebulan di mana pengguna mengakses situs web memiliki pengaruh yang cukup besar.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation

Pengguna potensial tinggi adalah mereka yang berusia kurang atau sama dengan 36 tahun, memiliki pendapatan area di atas atau sama dengan 384.864.700, menghabiskan lebih dari 65 menit di situs web, dan aktif menggunakan internet lebih dari 180 MB. Meskipun demikian, mereka menunjukkan minat yang lebih rendah terhadap iklan umum dengan tidak mengklik iklan (Clicked on Ad: No). Mereka merupakan target pasar yang sangat potensial untuk iklan, namun memerlukan pendekatan yang lebih personalisasi dan konten yang sangat relevan.

Kelompok pengguna engagement rendah terdiri dari pengguna yang lebih tua dari rata-rata diatas 36 tahun, memiliki pendapatan area yang lebih rendah dibawah 384.864.700, menghabiskan waktu yang lebih sedikit di situs kurang dari 65 menit), dan menggunakan internet secara keseluruhan dengan frekuensi yang lebih rendah kurang dari 180 MB. Meskipun mereka cenderung lebih terbuka terhadap iklan (Clicked on Ad: Yes), kualitas dan relevansi iklan sangat penting bagi mereka.

Untuk selengkapnya, dapat melihat jupyter notebook disini

Business Recommendation & Simulation

Memanfaatkan Ekstensi Iklan

Menulis Deskripsi yang Efektif

Membuat Copy Iklan yang Menarik

Membuat Posting dengan Gambar dan Video

Strategi Soft Selling

A/B Testing

Call to Action (CTA)

Monitoring dan Analisis Lanjutan

Untuk selengkapnya, dapat melihat jupyter notebook disini

Simulation

Asumsi:

- Marketing Cost per Customer = $1000
- Keuntungan dari Customer yang Mengklik Iklan = $1500
- Simulasi 500 customer
- Distribusi click on ad adalah 50:50

Tanpa machine learning

Jadi kita membagi distribusi sebagai berikut:

- Customer yang Klik Iklan = 250
- Customer yang Tidak Klik Iklan = 250

- **Marketing Cost**: Total marketing cost untuk 500 customer adalah 500 x $1000 = $500,000.
- **Revenue** hanya dihasilkan dari customer yang klik iklan. Jadi, dengan 250 customer yang klik iklan, revenue adalah 250 x $1500 = $375,000.
- **Profit** dihitung dengan mengurangkan total marketing cost dari total revenue. Jadi, profitnya adalah $375,000 - $500,000 = -$125,000.
- Dari simulasi tanpa penggunaan machine learning, perusahaan mengalami kerugian sebesar $125,000 dengan persentase kerugian 25%

Untuk selengkapnya, dapat melihat jupyter notebook disini

Simulation

Asumsi:

- Marketing Cost per Customer = $1000
- Keuntungan dari Customer yang Mengklik Iklan = $1500
- Simulasi 500 customer
- Distribusi click on ad adalah 50:50

Menggunakan machine learning

Jadi kita membagi distribusi sebagai berikut:

- Customer yang Klik Iklan = 258
- Customer yang Tidak Klik Iklan = 242

- **Marketing Cost**: Total marketing cost untuk 500 customer adalah 258 x $1000 = $258,000.
- **Revenue** hanya dihasilkan dari customer yang klik iklan. Jadi, dengan 250 customer yang klik iklan, revenue adalah 258 x $1500 = $387,000.
- **Profit** dihitung dengan mengurangkan total marketing cost dari total revenue. Jadi, profitnya adalah $387,000 - $258,000 = $129,000.
- Dari simulasi penggunaan machine learning, perusahaan mendapat keuntungan sebesar $129,000 dengan persentase keuntungan 50%

Untuk selengkapnya, dapat melihat jupyter notebook disini

Simulation

Dari Hasil Simulasi:

1. Dengan menggunakan machine learning dalam strategi pemasaran, perusahaan berhasil mengurangi total biaya pemasaran dari $500,000 menjadi $258,000. penggunaan machine learning telah membantu mengoptimalkan target pemasaran, sehingga jumlah orang yang menjadi target pemasaran dapat dikurangi dari 500 orang menjadi 258 orang.
2. Revenue yang dihasilkan dari customer yang mengklik iklan juga meningkat dari $375,000 menjadi $387,000.
3. Sebagai hasilnya, perusahaan mencapai keuntungan sebesar $129,000 dengan persentase keuntungan 50%, yang merupakan peningkatan signifikan dibandingkan dengan kerugian sebelumnya sebesar $125,000 dengan persentase kerugian 25% tanpa menggunakan machine learning.

Untuk selengkapnya, dapat melihat jupyter notebook disini