Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Putri Kirey Eki Yogaswari
Kirey.ey11@gmail.com
www.linkedin.com/in/putrikirey

I am a fresh graduate from Public Health with a focus on Health Statistics. During college, I have knowledge of situation analysis, designing health programs, leadership skills, systems thinking and data analysis. I have a career interest in data management and proficient in using SPSS and Ms. Excel. Has managed data up to 2000+.

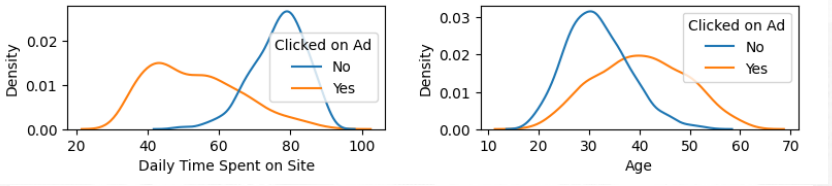
Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

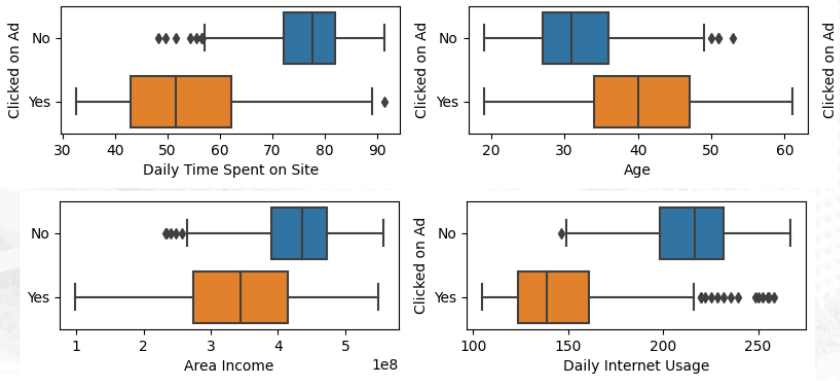
Overview

“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 ”

Customer Type and Behaviour Analysis on Advertisement



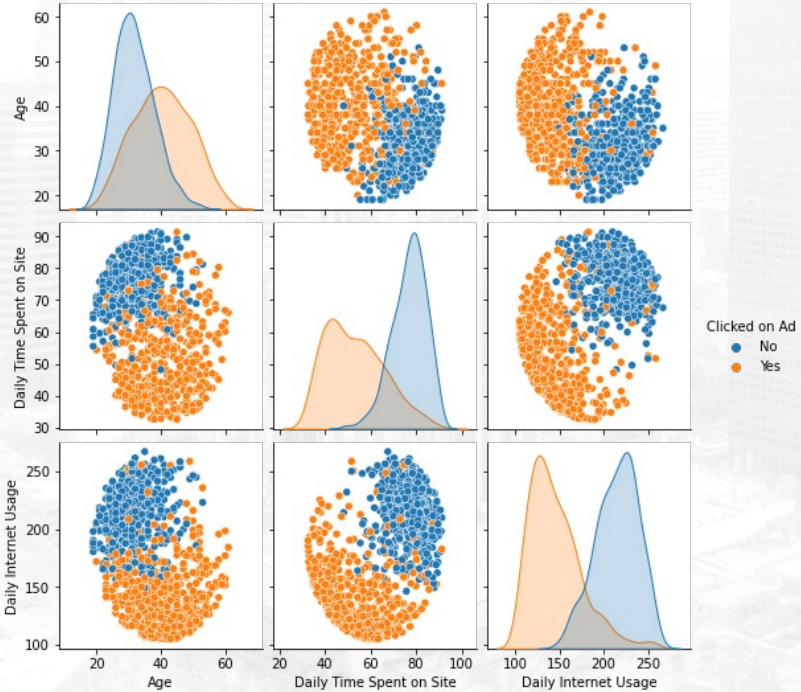


Observasi :

- Pengguna yang mengklik Ads adalah pengguna dengan Daily Time Spend on Site sekitar 40-45 menit. Sedangkan, pengguna yang tidak mengklik Ads adalah pengguna dengan Daily Time Spend on Site sekitar 75-80 menit.
- Pengguna yang mengklik Ads rata-rata ada pada usia(Age) 40 tahun. Sedangkan, pengguna yang tidak mengklik Ads sebagian besar ada pada usia(Age) 30 tahun.
- Pengguna dengan Daily Internet Usage sekitar 100-150 cenderung mengklik Ads. Sedangkan, pengguna dengan Daily Internet Usage sekitar 200-250 cenderung tidak mengklik Ads.
- Pengguna dengan Area Income yang lebih rendah cenderung mengklik Ads. sedangkan, pengguna dengan area income yang lebih besar cenderung tidak mengklik Ads.

Customer Type and Behaviour Analysis on Advertisement

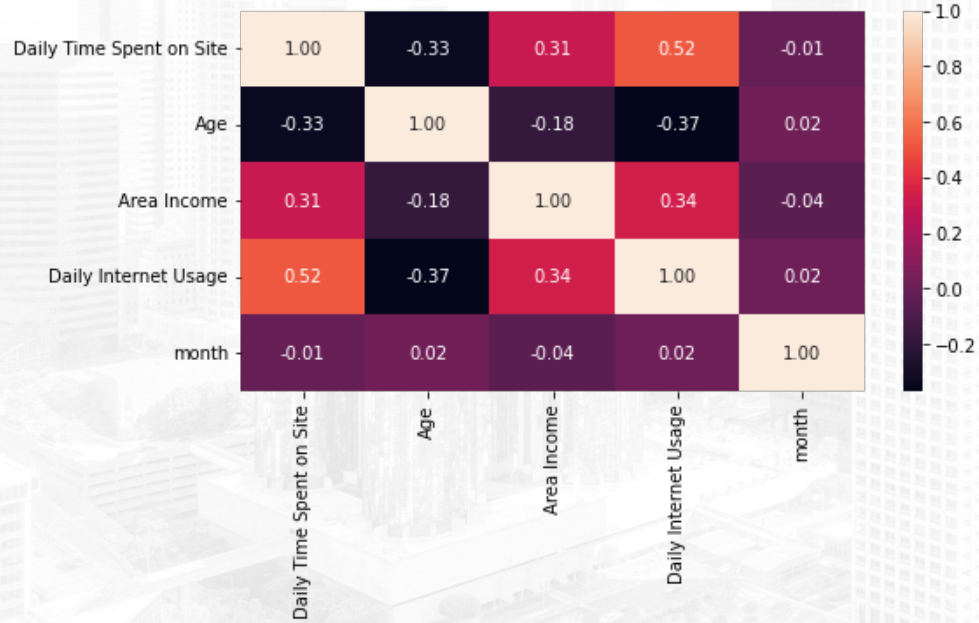
Bivariate Analysis



Observasi :

- Semakin tua usia (Age) user serta semakin sedikit Daily Internet Usage dan Daily Time Spent on Site maka seorang user cenderung mengklik Ads.
- Semakin sedikit Daily Internet Usage dan Daily Time Spent on Site maka seorang user cenderung mengklik Ads.

Multivariate Analysis



Observasi :

Dari korelasi di atas tidak ditemukan adanya multicorrelation (korelasi antar variable) sehingga kita dapat menggunakan semua feature untuk dilakukan modeling.

- Feature Daily Time Spent on Site berkorelasi positif cukup kuat dengan Daily Internet Usage.
- Feature Age berkorelasi negatif lemah dengan feature Daily Time Spent on Site, Area Income, dan Daily Internet Usage.
- Feature Area Income berkorelasi positif dengan feature Daily Time Spent on Site dan Daily Internet Usage dan berkorelasi negatif dengan feature Age.

Data Cleaning & Preprocessing

Rakamin Academy

Tahapan-tahapan yang dilakukan adalah:

Handling Missing Values

```
# imputation median to null values for numerical feature
df1.fillna(df1.median(),inplace=True)

#imputation mode for categorical feature
df1['Male'].fillna('Perempuan',inplace=True)

#recheck null values
df1.isnull().sum()
```

```
Daily Time Spent on Site    0
Age                         0
Area Income                 0
Daily Internet Usage        0
Male                        0
Timestamp                   0
Clicked on Ad               0
city                        0
province                    0
category                    0
month                       0
dtype: int64
```

Check Duplicated Data

```
[ ]  df1.duplicated().sum()
```

```
0
```

Extract Datetime Data

```
[ ]  import datetime
     df1.Timestamp = pd.to_datetime(df1.Timestamp)
     df1.Timestamp.dtype
```

```
dtype('<M8[ns]')
```

```
[ ]  df1['year']=df1.Timestamp.dt.year
     df1['month']=df1.Timestamp.dt.month
     df1['week']=df1.Timestamp.dt.isocalendar().week
     df1['day']=df1.Timestamp.dt.day
```

```
[ ]  #ubah tipe data feature week
     df1.week = df1.week.astype('int64')
```

Data Cleaning & Preprocessing

Rakamin Academy

Tahapan-tahapan yang dilakukan adalah:

Split Feature and Target

```
# Split features vs target
X = dfs[[col for col in dfs.columns if (str(dfs[col].dtype) != 'object') and col not in ['Clicked on Ad']]]
y = dfs['Clicked on Ad'].values
print(X.shape)
print(y.shape)
```

```
(1000, 33)
(1000,)
```

Feature Encoding

Label Encoding

```
[ ] #mengubah label Laki-Laki menjadi 1 dan label perempuan menjadi 0
    df1['Male'] = np.where(df1['Male']=='Laki-Laki',1,0)

    #check label
    df1.Male.unique()
```

```
array([0, 1])
```

```
[ ] df1['Clicked on Ad'].unique()
```

```
array(['No', 'Yes'], dtype=object)
```

```
[ ] #mengubah label Yes menjadi 1 dan label No menjadi 0
    df1['Clicked on Ad'] = np.where(df1['Clicked on Ad']=='Yes',1,0)

    #check label
    df1['Clicked on Ad'].unique()
```

```
array([0, 1])
```

Melakukan OHE pada feature category

```
[ ] ohe = ['province','category']
```

```
for cat in ohe :
  onehots = pd.get_dummies(df1[cat], prefix=cat)
  df1 = df1.join(onehots)
```

Data Modeling

	model_name	model	accuracy	recall	precision	duration
0	K-Nearest Neighbor	KNeighborsClassifier()	0.696667	0.640000	0.721805	0.010767
1	Logistic Regression	LogisticRegression()	0.500000	0.000000	0.000000	0.036223
2	Decision Tree	DecisionTreeClassifier()	0.946667	0.933333	0.958904	0.015433
3	Random Forest	(DecisionTreeClassifier(max_features='sqrt', r...	0.946667	0.933333	0.958904	0.997564
4	Gradient Boosting	([DecisionTreeRegressor(criterion='friedman_ms...	0.933333	0.913333	0.951389	0.652749

Berdasarkan proses pemodelan, model dengan skor akurasi tertinggi adalah Decision tree, Random forest dan Gradient boosting. Namun jika dilihat dari durasinya, decision tree merupakan model dengan durasi waktu paling rendah. Ini adalah hasil pemodelan tanpa menangani normalisasi data dan skor ini dapat ditingkatkan lebih lanjut melalui penerapan normalisasi pada kumpulan data.

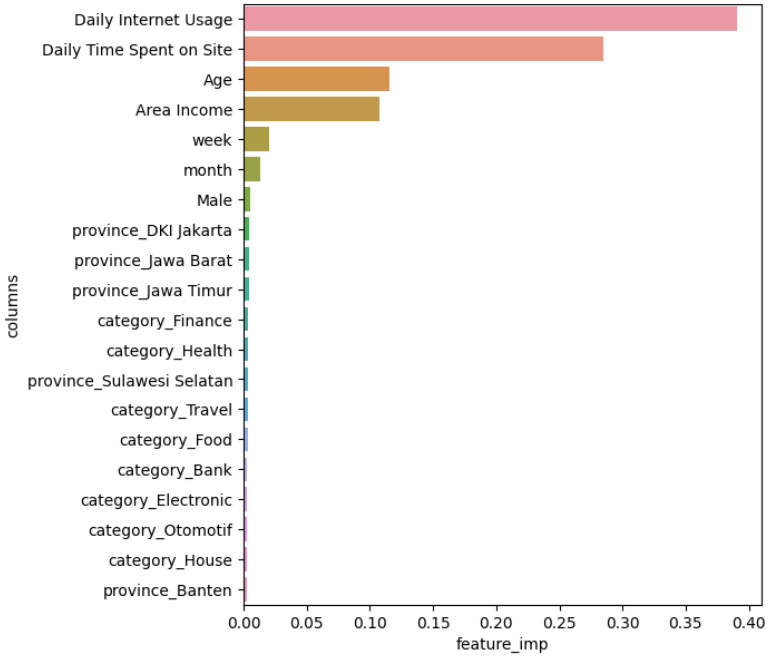
Setelah data di normalisasi

	model_name	model	accuracy	recall	precision	duration
0	K-Nearest Neighbor	KNeighborsClassifier()	0.786667	0.740000	0.816176	0.002395
1	Logistic Regression	LogisticRegression()	0.936667	0.900000	0.971223	0.014669
2	Decision Tree	DecisionTreeClassifier()	0.936667	0.926667	0.945578	0.005029
3	Random Forest	(DecisionTreeClassifier(max_features='sqrt', r...	0.943333	0.926667	0.958621	0.308573
4	Gradient Boosting	([DecisionTreeRegressor(criterion='friedman_ms...	0.930000	0.913333	0.944828	0.332312

Akurasinya mengalami peningkatan untuk semua model khususnya pada KNN dan Logistic Regression yang mengalami peningkatan signifikan dibandingkan hasil sebelumnya. Tujuan model kami adalah memprediksi jumlah maksimum calon pelanggan yang mengklik iklan. Oleh karena itu, kita harus meminimalisir False Positive dimana pelanggan yang tidak mengklik iklan diprediksi akan salah mengklik iklan. Hal ini akan menyebabkan terjadinya retargeting pasar yang salah dan berpotensi menimbulkan kerugian karena kita telah mengeluarkan biaya pemasaran pada target yang salah.

Oleh karena itu kita harus mengoptimalkan skor precision dengan tetap mempertimbangkan metriks lain agar skornya maksimal.
Terakhir, model random forest dipilih dengan mempertimbangkan skor precision dan accuracy yang tinggi.

Feature Importance



Berdasarkan grafik tingkat kepentingan fitur di atas terlihat ada 4 fitur yang memiliki korelasi tinggi dalam membangun model, yaitu :

- Penggunaan Internet Sehari-hari
- Waktu Sehari-hari yang Dihabiskan di Situs
- Usia
- Pendapatan Daerah

Fitur dengan korelasi tertinggi dalam memprediksi iklan yang diklik pelanggan adalah Penggunaan Internet Harian dan Waktu yang Dihabiskan di Situs setiap hari. Hal ini dibuktikan berdasarkan analisis EDA sebelumnya bahwa pelanggan dengan penggunaan internet harian yang lebih rendah dan waktu yang dihabiskan di situs memiliki potensi mengklik iklan yang lebih tinggi.

Ciri-ciri lain yang juga mempunyai tingkat kepentingan yang tinggi adalah umur dan pendapatan daerah. Hal ini didukung dengan pemetaan sebaran pelanggan berdasarkan usia dan pendapatan daerah menunjukkan bahwa pelanggan yang memiliki usia lebih muda dan pendapatan lebih tinggi memiliki potensi mengklik iklan yang rendah.

Rekomendasi Bisnis

1. Personalisasi Iklan Berdasarkan Waktu yang Dihabiskan di Situs (Daily Time Spend on Site):
Untuk pengguna dengan Daily Time Spend on Site sekitar 40-45 menit, kami dapat mengoptimalkan iklan agar lebih menarik bagi mereka. Mungkin kami bisa menargetkan iklan yang lebih interaktif atau konten yang lebih singkat dan tajam.
Bagi pengguna dengan Daily Time Spend on Site sekitar 75-80 menit, kami bisa mencoba menyajikan iklan yang lebih informatif atau menarik perhatian mereka dengan promosi khusus.

2. Optimalkan Berdasarkan Daily Internet Usage:
Pelanggan dengan Daily Internet Usage sekitar 100-150 cenderung mengklik iklan, jadi fokuskan upaya pemasaran pada kelompok ini. Kami bisa mengirim iklan yang lebih sesuai dengan minat mereka.
Untuk pengguna dengan Daily Internet Usage sekitar 200-250 yang cenderung tidak mengklik iklan, kami mungkin perlu menyajikan iklan yang lebih menarik atau mengurangi frekuensi tampilan iklan kepada mereka.

3. Sesuaikan Target Demografi Berdasarkan Usia (Age):
Jika pengguna dengan usia rata-rata 40 tahun lebih cenderung mengklik iklan, kami bisa lebih fokus pada penargetan kelompok usia yang lebih tua. Kami dapat membuat iklan yang lebih relevan dengan preferensi dan kebutuhan mereka.
Untuk pengguna sebagian besar berusia 30 tahun yang tidak cenderung mengklik iklan, kami bisa mempertimbangkan strategi khusus untuk menarik perhatian mereka, seperti penawaran eksklusif atau konten yang lebih menarik bagi kelompok usia ini.

4. Maksimalkan Penggunaan Data:
Gunakan data usia dan pendapatan daerah untuk mengidentifikasi dan menargetkan kelompok pelanggan yang memiliki potensi tinggi untuk mengklik iklan. Ini dapat membantu kami mengalokasikan anggaran pemasaran dengan lebih efisien.

Simulasi

Asumsi :

- Marketing cost per customer = Rp 10.000
- Profit gained per customer who clicked on ads = Rp 15.000
- Kami akan melakukan simulasi dengan sasaran populasi 500 customers

```
df['Clicked on Ad'].value_counts()
```

```
No     500
Yes    500
Name: Clicked on Ad, dtype: int64
```

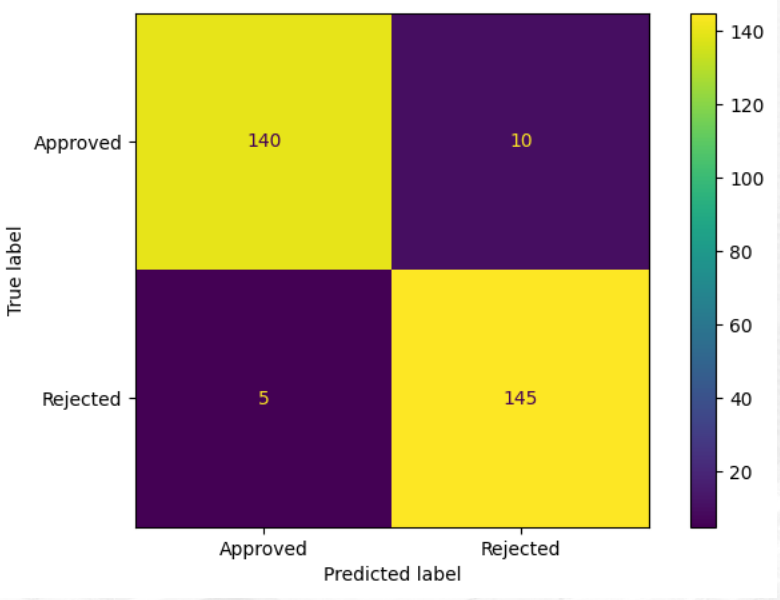
Customers dibagi menjadi dua kelompok dengan probabilitas 50:50. Ada customer yang mengklik iklan (250 customer) dan customer yang tidak mengklik iklan (250 customer).

Marketing cost = 500 customers X Rp 10.000 = Rp 5.000.000
Conversion rate = 250/500 *100% = 50%
Revenue = 250 customers X Rp 15.000 = Rp 3.750.000
Profit = Revenue - Cost = - Rp 1.250.000

Berdasarkan simulasi di atas, besarnya potensi kerugian sebesar Rp 1.250.000 dengan rate sekitar 25%

Business Recommendation & Simulation

Dengan Menggunakan Machine Learning Model



Berdasarkan confussion metrix, kita dapat mengelompokkan pelanggan berdasarkan prediksi siapa yang mengklik iklan (145 pelanggan) dan siapa yang tidak mengklik iklan (155 pelanggan).

Marketing cost = 145 customers X Rp 10.000 = Rp 1.450.000
Conversion rate = 140/145 *100% = 96,55%
Revenue = 140 customers X Rp 15.000 = Rp 2.100.000
Profit = Revenue - Cost = Rp 650.000

Berdasarkan simulasi di atas, besarnya keuntungan sebesar Rp 650.000 dengan tingkat keuntungan 46,55%.

Rakamin Academy

Terima Kasih!