Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Citta Mudita
citmud10@gmail.com
https://www.linkedin.com/in/citta-mudita-40b91b22b/

I have a strong interest in data analysis and data science. My experience in digital marketing has made me an expert in conducting exploratory data analysis and creating data visualizations to address business challenges. I have received awards as the Top 2 Student in the Data Science Bootcamp and as the Most Outstanding Student for the Final Project. Additionally, my team was recognized as the Best Final Project Team at Data Science Bootcamp.

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 ”

Rakamin Academy

Exploration Data Analysis (EDA)

Statistical analysis

Featur Numeric

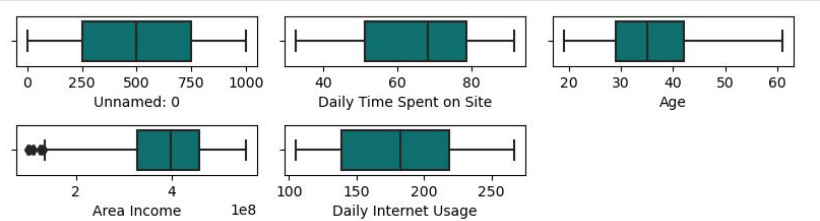
	Unnamed: 0	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage
count	1000.000000	987.000000	1000.000000	9.870000e+02	989.000000
mean	499.500000	64.929524	36.009000	3.848647e+08	179.863620
std	288.819436	15.844699	8.785562	9.407999e+07	43.870142
min	0.000000	32.600000	19.000000	9.797550e+07	104.780000
25%	249.750000	51.270000	29.000000	3.286330e+08	138.710000
50%	499.500000	68.110000	35.000000	3.990683e+08	182.650000
75%	749.250000	78.460000	42.000000	4.583554e+08	218.790000
max	999.000000	91.430000	61.000000	5.563936e+08	267.010000

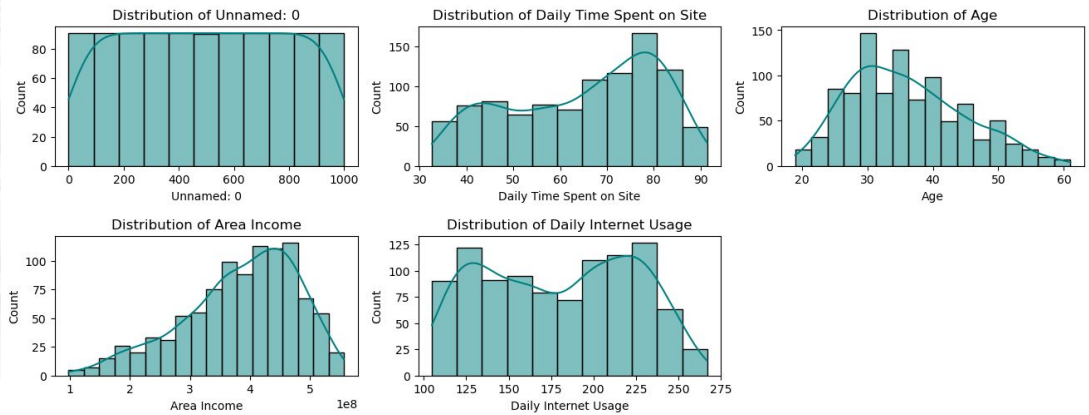
Featur Categoric

	count	unique	top	freq
Male	997	2	Perempuan	518
Clicked on Ad	1000	2	No	500
city	1000	30	Surabaya	64
province	1000	16	Daerah Khusus Ibukota Jakarta	253
Timestamp	1000	997	5/26/2016 15:40	2

Visualisasi Data

Univariate Analysis





Fitur Area Income memiliki **distribusi negative skewed**, artinya kebanyakan customer berpendapatan rendah dibandingkan berpendapatan tinggi.

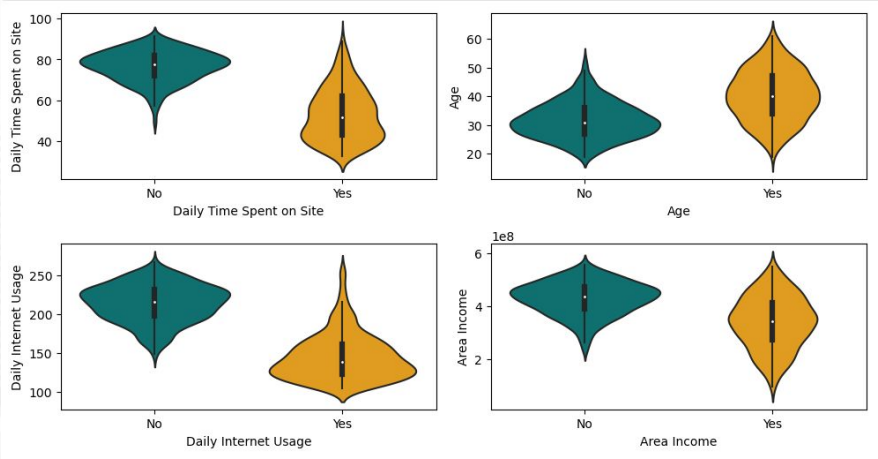
Fitur Age memiliki **distribusi positive skewed,** artinya kebanyakan customer memiliki umur lebih dari 30 tahun dibandingkan customer yang memiliki umur kurang dari itu.

Feature yang memiliki **distribusi normal** : Daily Time Spent on Site and Daily Internet Usage

Feature yang memiliki **outlier** : Area Income

Overview

Univariate Analysis



Daily Time Spent on Site:
Waktu yang dihabiskan oleh user yang mengklik iklan memiliki distribusi skew kanan artinya waktu yang dihabiskan cenderung sedikit yaitu 40-45 menit.
Sedangkan waktu yang dihabiskan user yang tidak menklik iklan memiliki distribusi skew kiri artinya waktu yang di habiskan cenderung banyak yaitu 75-80 menit.

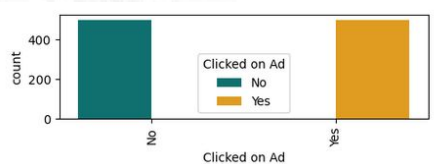
Usia Pengguna:
Usia yang mengklik iklan memiliki distribusi normal dengan rata-ratanya adalah 40 tahun. Sedangkan usia user yang tidak mengklik iklan memiliki distribusi skew kanan artinya usian pengguna cenderung lebih muda yaitu di bawah 40 tahun.

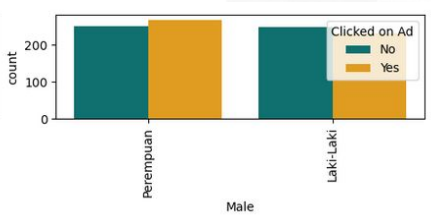
Daily Internet Usage:
Penggunaan internet harian untuk user yang mengklik iklan memiliki distribusi skew kanan, artinya penggunaan internet harian user cenderung rendah yaitu 100-150 menit. Sedangkan penggunaan internet harian untuk user yang tidak mengklik iklan memiliki disribusi skew kiri, artinya penggunaan internet harian user cenderung tinggi yaitu 200-250 menit.

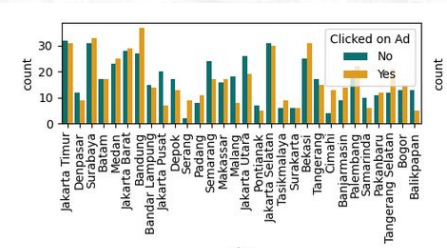
Area Income
Pendapatan wilayah geografis user yang mengklik iklan memiliki distribusi normal, sedangkan pendapatan wilayah geografis user yang tidak mengklik iklan memiliki distribusi skew kiri yang artinya user dengan pendapatan wilayah geografis tinggi.

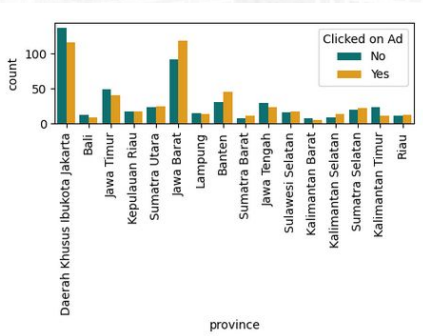
Overview

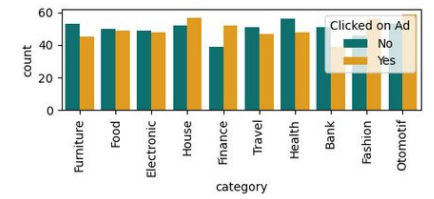
Univariate Analysis











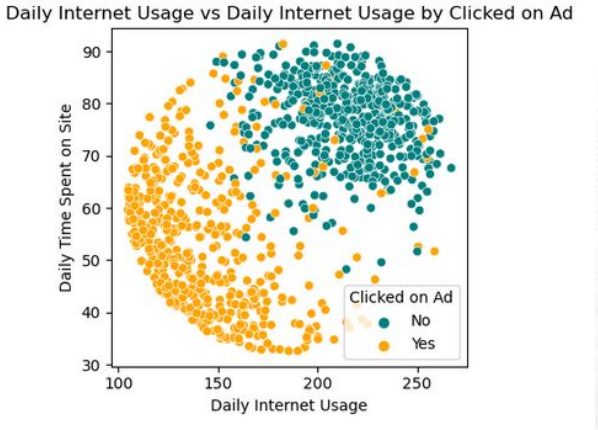
Jumlah "Yes" dan "No" pada fitur "Clicked on Ad" seimbang. Keseimbangan ini mengindikasikan bahwa dataset memiliki distribusi yang relatif seragam antara pengguna yang mengklik iklan ("Yes") dan yang tidak ("No").

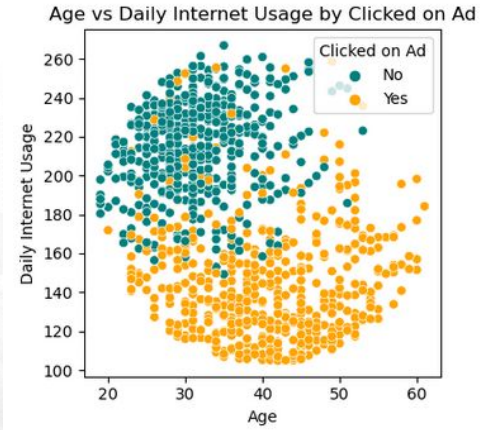
Jumlah perempuan dan laki-laki dalam fitur "Male" tidak terlalu timpang. Ini menunjukkan bahwa dataset memiliki seimbang antara kedua jenis kelamin, yang bisa berguna dalam analisis selanjutnya.

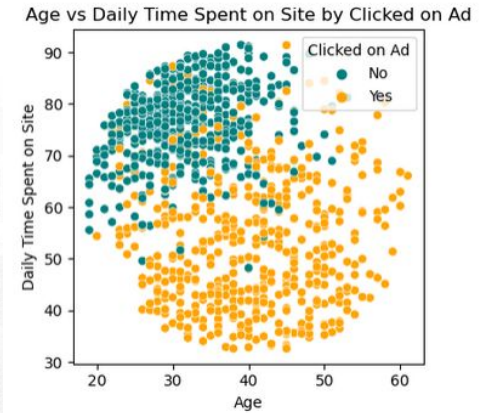
Fitur "Province" didominasi oleh dua nilai utama, yaitu "DKI Jakarta" dan "Jawa Barat". Hal ini menunjukkan bahwa sebagian besar pengguna berasal dari dua provinsi ini, sementara provinsi-provinsi lainnya mungkin memiliki kontribusi yang lebih rendah dalam dataset.

Overview

Bivariate Analysis







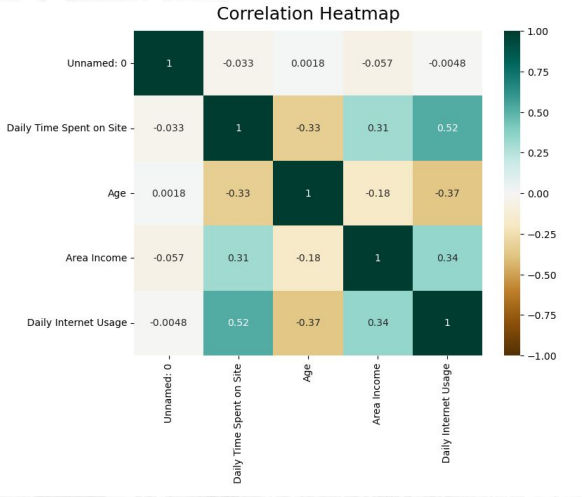
User dengan penggunaan internet harian rendah memiliki waktu yang dihabiskan disitus juga rendah yang memiliki kecenderungan untuk mengklik iklan, sedangkan user dengan penggunaan internet harian tinggi memiliki waktu yang dihabiskan di situs juga tinggi memiliki kecenderungna untuk tidak mengklik iklan.

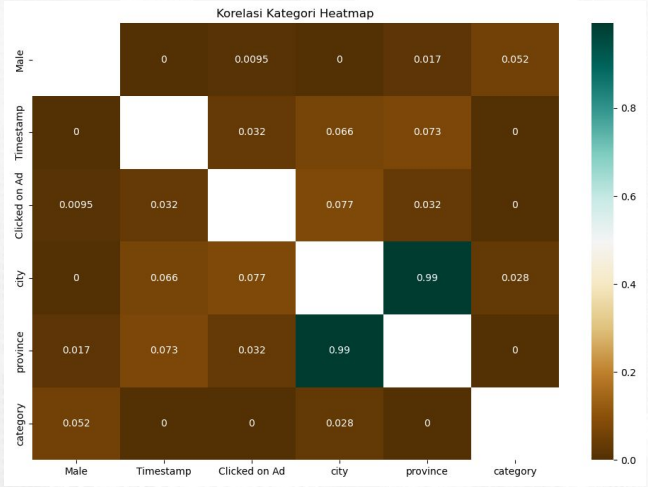
User dengan usia yang lebih tua, penggunaan internet harian yang lebih rendah, dan waktu yang dihabiskan di situs yang lebih rendah cenderung untuk mengklik iklan.

User dengan usia yang lebih muda, penggunaan internet harian tinggi, dan waktu yang dihabiskan di situs juga tinggi memiliki kecenderungan untuk tidak mengklik iklan.

Overview

Multivariate analysis





Dari hasil korelasi yang diperoleh melalui heatmap, ditemukan fitur yang memiliki korelasi yang kuat (redundan) yaitu fitur city dengan province. Oleh karena itu, fitur city dan province tidak digunakan dalam pemodelan.

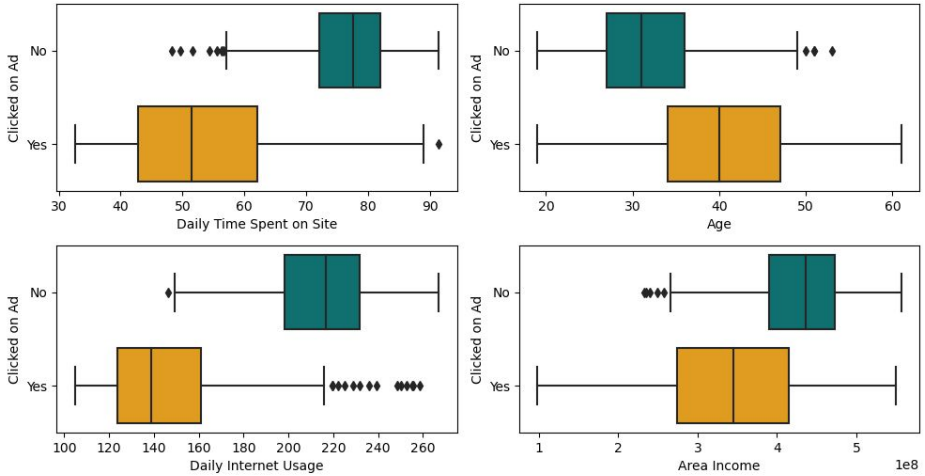
Penggunaan internet harian memiliki korelasi positif yang tinggi (51%) dengan waktu harian yang dihabiskan di website. Artinya, semakin sering pengguna menggunakan internet, semakin lama mereka menghabiskan waktu di situs web.

Usia pengguna memiliki korelasi negatif dengan tiga fitur lainnya: pendapatan rata-rata geografis pengguna, waktu yang dihabiskan di website, dan penggunaan internet harian. Korelasi tertinggi adalah dengan penggunaan internet harian (37%). Hal ini menunjukkan bahwa semakin tua usia pengguna, semakin jarang mereka menggunakan internet harian.

Source code

Univariate terhadap target

[81]:
```
# bivariate analysis: boxplot
colors = sns.color_palette(["teal", "orange"])
plt.figure(figsize = (10,10))
for i in range(len(num_2)):
  plt.subplot(4, 2, i + 1)
  sns.boxplot(x = df[num_2[i]], y = df['Clicked on Ad'], palette=colors)
  plt.tight_layout()
```



Data Cleaning & Preprocessing

Handle Missing Value

```
Daily Time Spent on Site    13
Area Income                 13
Daily Internet Usage        11
Male                         3
Unnamed: 0                   0
Age                          0
Timestamp                    0
Clicked on Ad                0
dtype: int64
```

Mengisi nilai null menggunakan median (numeric) dan mode (categoric)

Handle Duplicate Value

```
df_1.duplicated().sum()
```

```
0
```

Feature Extraction

Mengekstraksi feature timestamp menjadi tahun, bulan, pekan, dan hari.

Feature Encoding

Label Encoding
- Male
- Clicked on Ad

One Hot Encoding
- Category

Feature Selection

Drop feature
- Unnamed 0
- City
- Province
- Timestamp

Handle Outlier

Tidak melakukan handle outlier karna menggunakan model yang robust terhadap outlier

Split Data Train and Data Test

Data Train

70%

Data Test

30%

Modeling

- Without Standardization
- With Standardization

Data Cleaning & Preprocessing

Rakamin Academy

Handling Missing Value

[31]:
```
missing_values=df_1.isnull().sum().sort_values(ascending=False)[:8]
missing_values
```

[31]:
```
Daily Time Spent on Site    13
Area Income                 13
Daily Internet Usage        11
Male                         3
Unnamed: 0                   0
Age                          0
Timestamp                    0
Clicked on Ad                0
dtype: int64
```

[32]:
```
total_rows = len(df_1)
missing_percentage = (missing_values / total_rows) * 100
missing_percentage
```

[32]:
```
Daily Time Spent on Site    1.3
Area Income                 1.3
Daily Internet Usage        1.1
Male                        0.3
Unnamed: 0                  0.0
Age                         0.0
Timestamp                   0.0
Clicked on Ad               0.0
dtype: float64
```

Feature dengan missing value

- Daily Time Spent on Site sebanyak 13 data (1.3%)
- Area Income sebanyak 13 data (1.3%)
- Daily Internet Usage sebanyak 11 data (1.1%)
- Male sebanyak 3 data (0.3%).

[33]:
```
# Mengisi nilai kosong dengan median pada feature numerik dan modus pada feature kategorik
df_1['Daily Time Spent on Site'].fillna(df_1['Daily Time Spent on Site'].median(),
                                        inplace = True)
df_1['Area Income'].fillna(df_1['Area Income'].median(),
                           inplace = True)
df_1['Daily Internet Usage'].fillna(df_1['Daily Internet Usage'].median(),
                                    inplace = True)
df_1['Male'].fillna(df_1['Male'].mode()[0], inplace=True)
```

Data Modeling

Hasil Without Standardization

model	Recall_test	Recall_train	Accuracy_test	Accuracy_train
Logistic Regression	0.927928	0.966581	0.940	0.97625
Decision Tree	0.918919	0.946015	0.915	0.95000
Random Forest	0.963964	0.964010	0.940	0.97750
Adaboost	0.927928	0.979434	0.945	0.99000
Gradient Boosting	0.936937	0.974293	0.925	0.98500

- Sebagian besar model menunjukkan hasil kinerja yang serupa antara data uji dan data latih, yang menandakan bahwa tidak terdapat perbedaan yang signifikan yang mengindikasikan overfitting atau underfitting
- Dari hasil pemodelan pada data uji, nilai recall lebih tinggi pada model Random Forest dan Gradient Boosting.
- Selain akurasi, hasil recall pada data uji juga menunjukkan hasil yang lebih baik pada model Gradient Boosting.

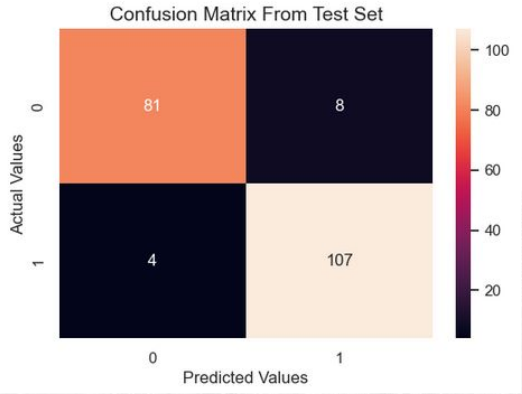
Hasil With Standardization

model	Recall_test	Recall_train	Accuracy_test	Accuracy_train
Logistic Regression	0.954955	0.969152	0.955	0.97625
Decision Tree	0.918919	0.946015	0.915	0.95000
Random Forest	0.963964	0.964010	0.940	0.97750
Adaboost	0.927928	0.979434	0.945	0.99000
Gradient Boosting	0.936937	0.976864	0.925	0.98625

- Terjadi peningkatan pada model, hal nni menunjukkan bahwa penggunaan Min-Max Scaler telah memberikan dampak positif pada performa model.
- Terjadi perubahan dalam hasil kinerja terbaik, di mana hasil akurasi tes tertinggi ditemukan pada model Random Forest dan Gradient Boosting. Hal ini menunjukkan bahwa kedua model tersebut adalah yang terbaik dalam memprediksi data uji setelah dilakukan normalisasi/standarisasi. Di pilih model Random Forest karena nilai recal antara train dan test gapnya lebih dekat dari pada yang Gradient Boosting.

Data Modeling

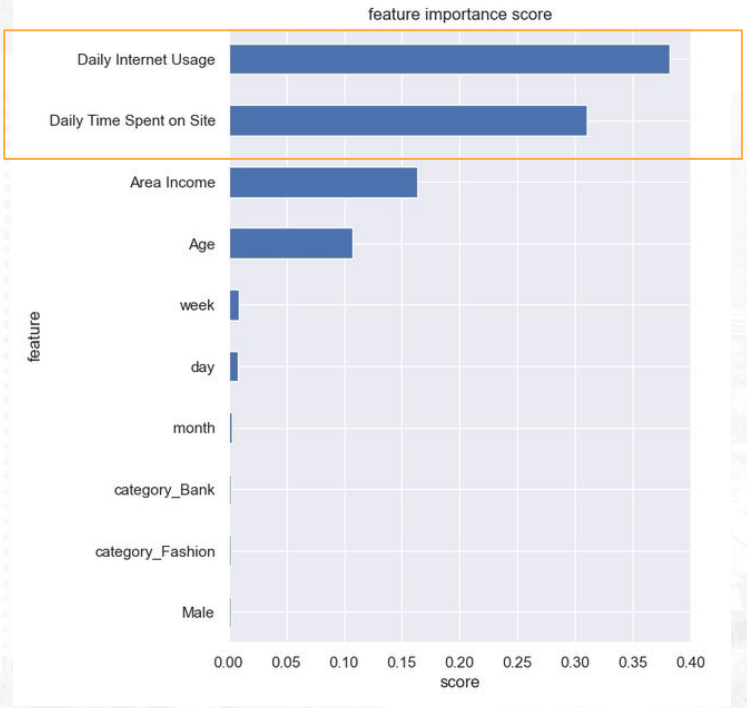
Confusion matrix



Recall mengukur rasion antara jumlah user yang diprediksi mengklik iklan dan benar-benar mengklik iklan. dengan jumlah user yang diprediksi tidak mengklik iklan tetapi ada kenyatannya mengklik iklan. Memaksimalkan recall ini akan meminimalkan jumlah user yang salah prediksi akan mengklik iklan.

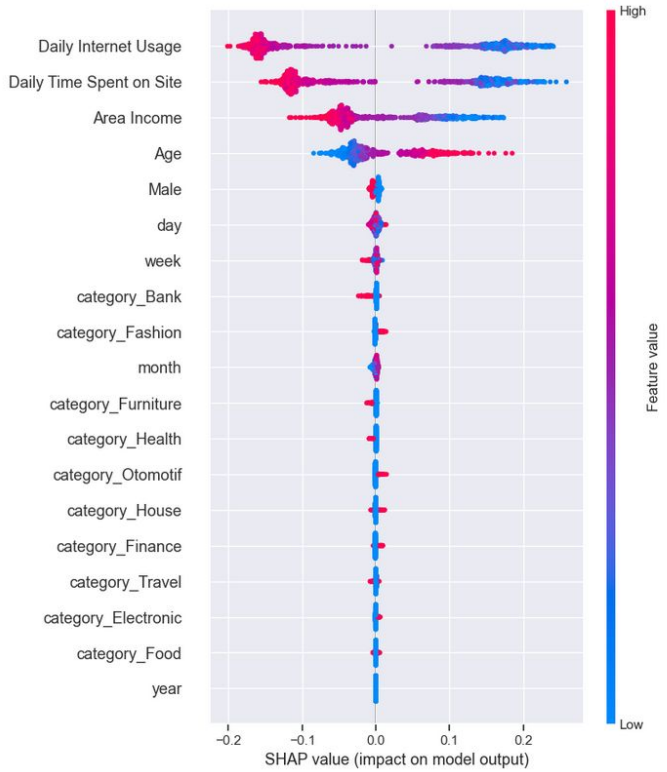
Berdasarkan model yagn digunakan, terdapat 2 feature yang paling berpengaruh yaitu Daily Internet Usage dan Dailt Time Spent on Site. Terdapat feature yang cukup berpengaruh yaitu feature Area Income

Feature Important



Business Recommendation & Simulation

Feature Important



Berdasarkan model yang digunakan, terdapat 2 feature yang paling berpengaruh yaitu Daily Internet Usage dan Dailt Time Spent on Site.

Semakin rendah Daily Internet Usage, maka semakin banyak user mengklik iklan, sedangkan semakin tinggi Daily Internet Usage, maka semakin sedikit user mengklik iklan.

Begitu pun dengan Daily Time Spent on Site. Semakin rendah Daily Time Spent on Site, maka semakin banyak user mengklik iklan, sedangkan semakin tinggi Daily Time Spent on Site, maka semakin sedikit user mengklik iklan.

Selain itu terdapat feature yang lain yang mempengaruhi yaitu Area Income, semaking tinggi area income, semakin tinggi area income user, semakin cenderung untuk tidak mengklik iklan.

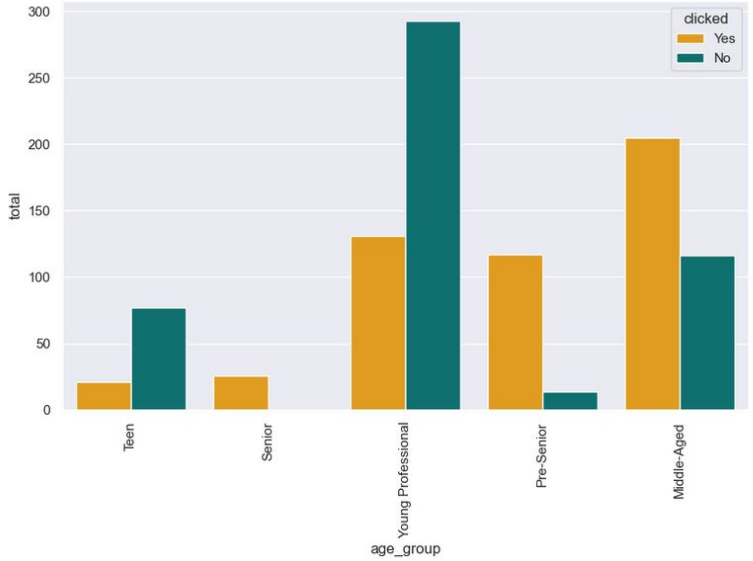
Business Recommendation & Simulation

Interpretasi

Hubungan antara usia user dengan mengklik iklan

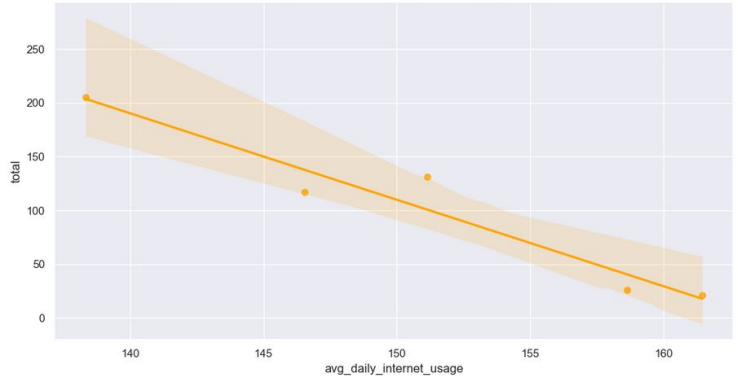
User dengan usia di atas 35 tahun yaitu Middle Age, Pre Senior dan Senior lebih banyak mengklik iklan

User dengan usia di bawah 35 tahun yaitu Young Profesional, teen lebih sedikit mengklik iklan



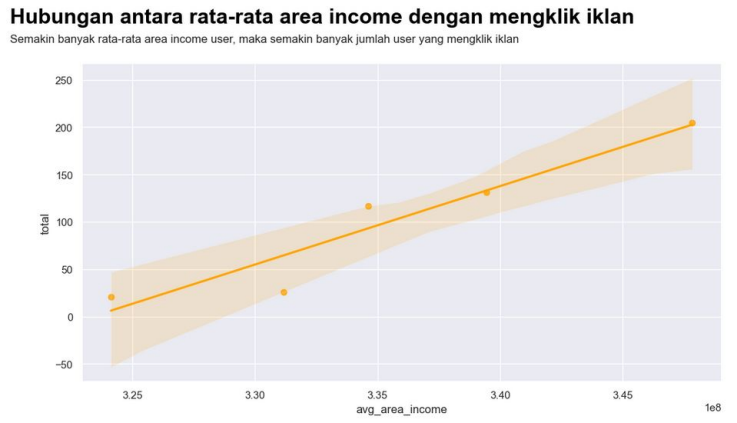
Hubungan antara rata-rata pengunaan internet harian dengan mengklik iklan

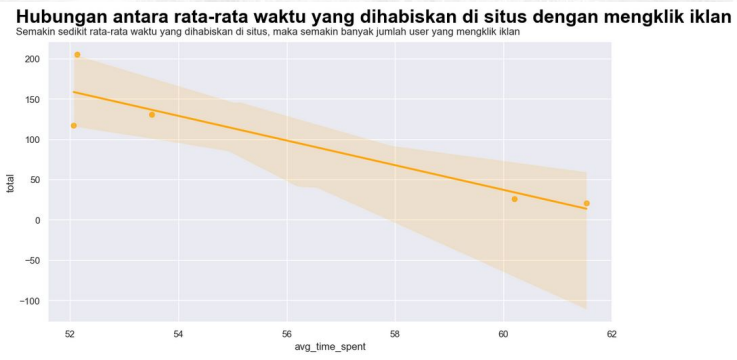
Semakin sedikit rata-rata penggunaan internet harian, maka semakin banyak jumlah user yang mengklik iklan



Business Recommendation & Simulation

Interpretasi





Dari data yang diperoleh, terdapat 2 segmen user, yaitu segmen user aktif dan non aktif.
User aktif memiliki karakteristik dengan usia di bawah 35 tahun yaitu masuk dalam grup teen dan young profesional, memiliki penggunaan internet yang tinggi, serta banyak menghabiskan waktu di situs, memiliki pendapatan yang lebih rendah.

Sementara user non aktif memiliki karakteristik usia di atas 35 tahun yaitu grup Middle Age, Pre Senior dan Senior, memiliki penggunaan internet rendah yang mempengaruhi waktu yang dihabiskan di situs juga rendah, serta memiliki pendapatan yang lebih tinggi. User non-aktif cenderung untuk mengklik iklan dibandingkan dengan user aktif.

Business Recommendation & Simulation

Recommendation Business

Penyesuaian Strategi Periklanan

- Untuk pengguna aktif (usia di bawah 35 tahun), perusahaan dapat fokus pada iklan yang menarik bagi mereka, seperti penawaran khusus, promosi, atau iklan yang lebih interaktif.
- Untuk pengguna non-aktif (usia di atas 35 tahun), perusahaan dapat menggunakan pendekatan iklan yang lebih informatif, menekankan kualitas produk atau layanan, dan menawarkan solusi yang relevan.

Pengelolaan Iklan

- Mengelola frekuensi dan jumlah iklan yang ditampilkan kepada pengguna aktif. Terlalu banyak iklan dapat membuat mereka merasa terganggu dan mengurangi respons positif terhadap iklan.

Business Recommendation & Simulation

Rakamin Academy

Simulation

COST = Cost Ads/User IDR 1000 X Total User 1000

CVR before Model = 50%

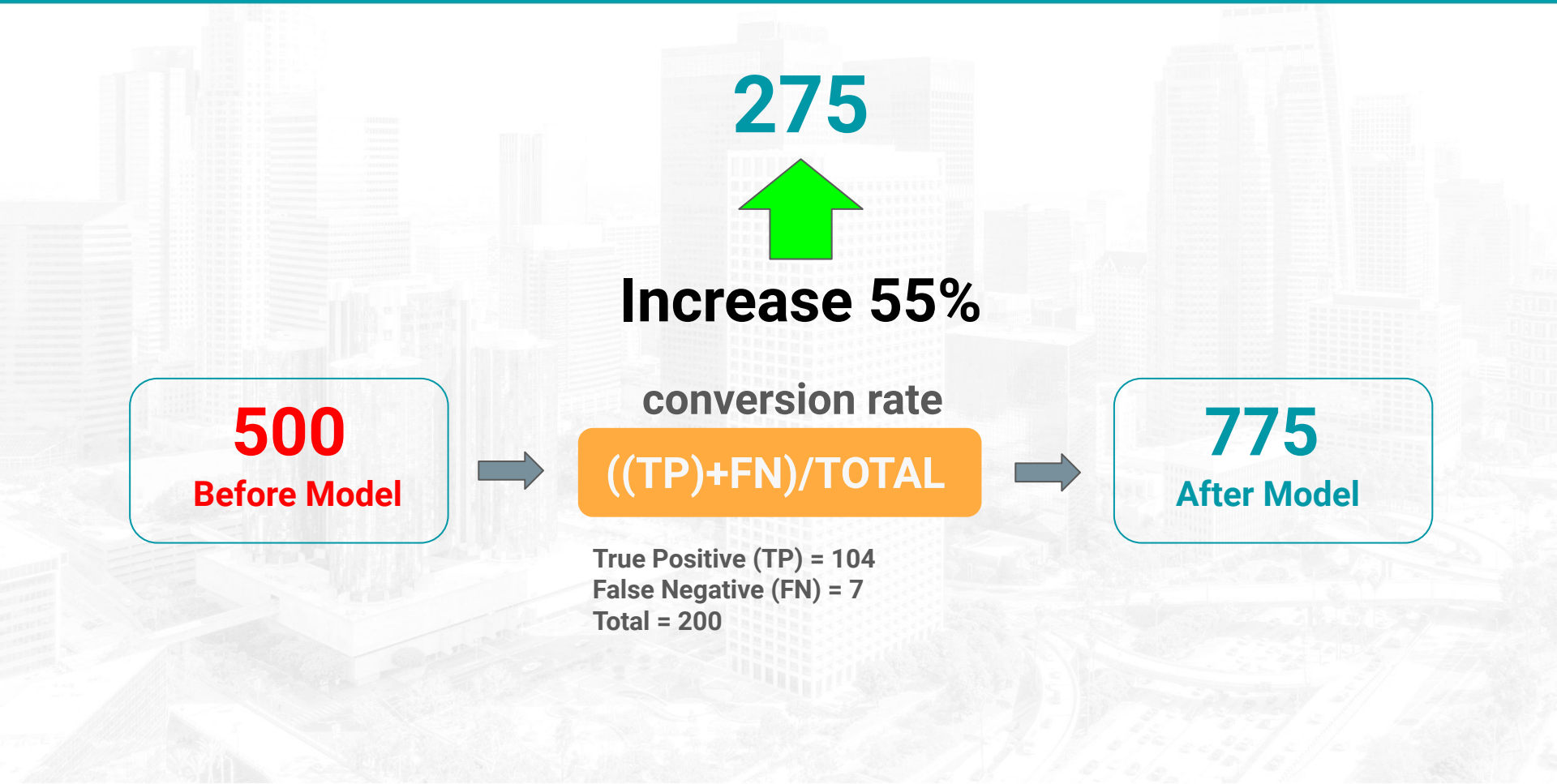
COST = IDR 1.000.000

RV = IDR 5000

PT = REVENUE - COST

Keterangan
RV = Revenue
PT = Profit

Business Recommendation & Simulation



Business Recommendation & Simulation

Rakamin Academy

cvr 50%

500
Before Model

RE
IDR 2.500.000

PT
IDR 1.500.000

27,5%

275
Increase

RE
IDR 1.350.000

Increase profit
1,9 X

cvr 77,5%

775
After Model

RE
IDR 3.850.000

PT
IDR 2.850.000

Untuk selengkapnya, dapat melihat jupyter notebook disini