Predict Customer
Clicked Ads Classification by Using Machine Learning

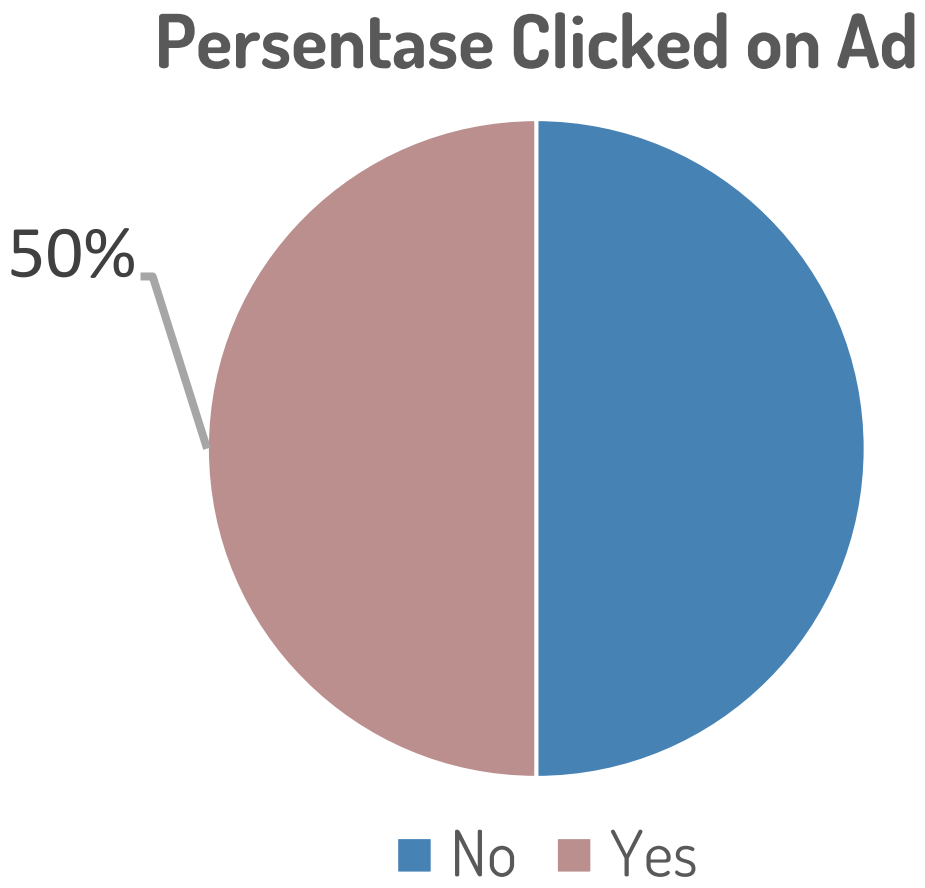
Oleh : Friska Yuliantika S

Link GitHub Repo

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

Link GitHub Repo

Statistical analysis

Numeric Variable

	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Day
count	987.00	1000	987.00	989.00	1000
mean	64.93	36.01	384,864,670.64	179.86	15.48
std	15.84	8.79	94,079,989.57	43.87	8.73
min	32.60	19	97,975,500.00	104.78	1
25%	51.27	29	328,632,990.00	138.71	8
50%	68.11	35	399,068,320.00	182.65	15
75%	78.46	42	458,355,450.00	218.79	23
max	91.43	61	556,393,600.00	267.01	31

Datetimestamp Variable

	Timestamp
count	1000
unique	997
top	2016-05-26 15:40:00
freq	2
first	2016-01-01 02:52:00
last	2016-07-24 00:22:00

Categoric Variable

	Gender	Clicked on Ad	city	province	category	Month	Day_Name
count	997	1000	1000	1000	1000	1000	1000
unique	2	2	30	16	10	7	7
top	Perempuan	No	Surabaya	DKI Jakarta	Otomotif	February	Sunday
freq	518	500	64	253	112	160	159

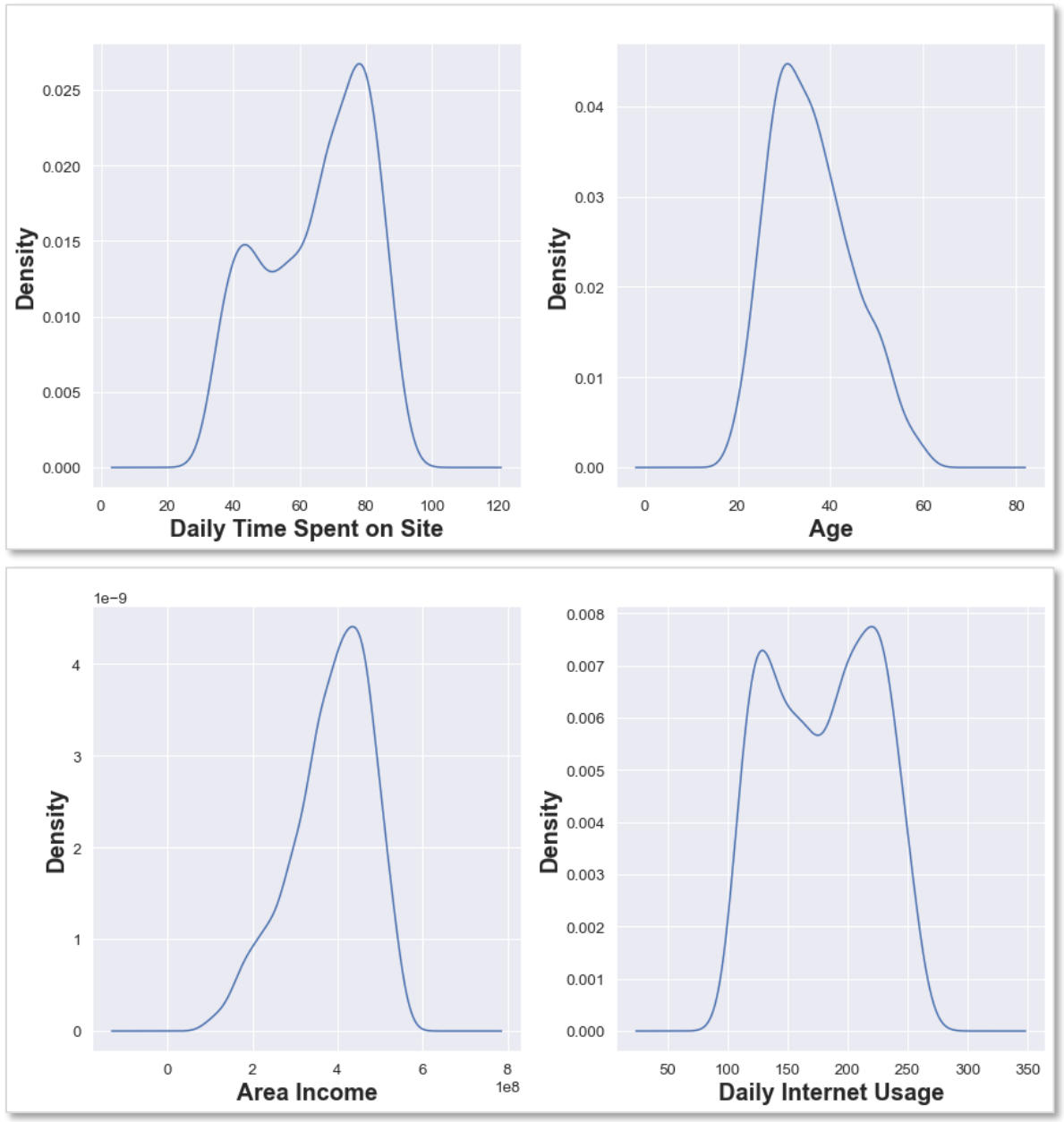
Data Cleansing

- Melakukan feature engineering pada variable 'Timestamp'
- Merubah tipe data dan memisahkan sesuai dengan tipenya
- Merubah nama kolom 'Male' menjadi Gender

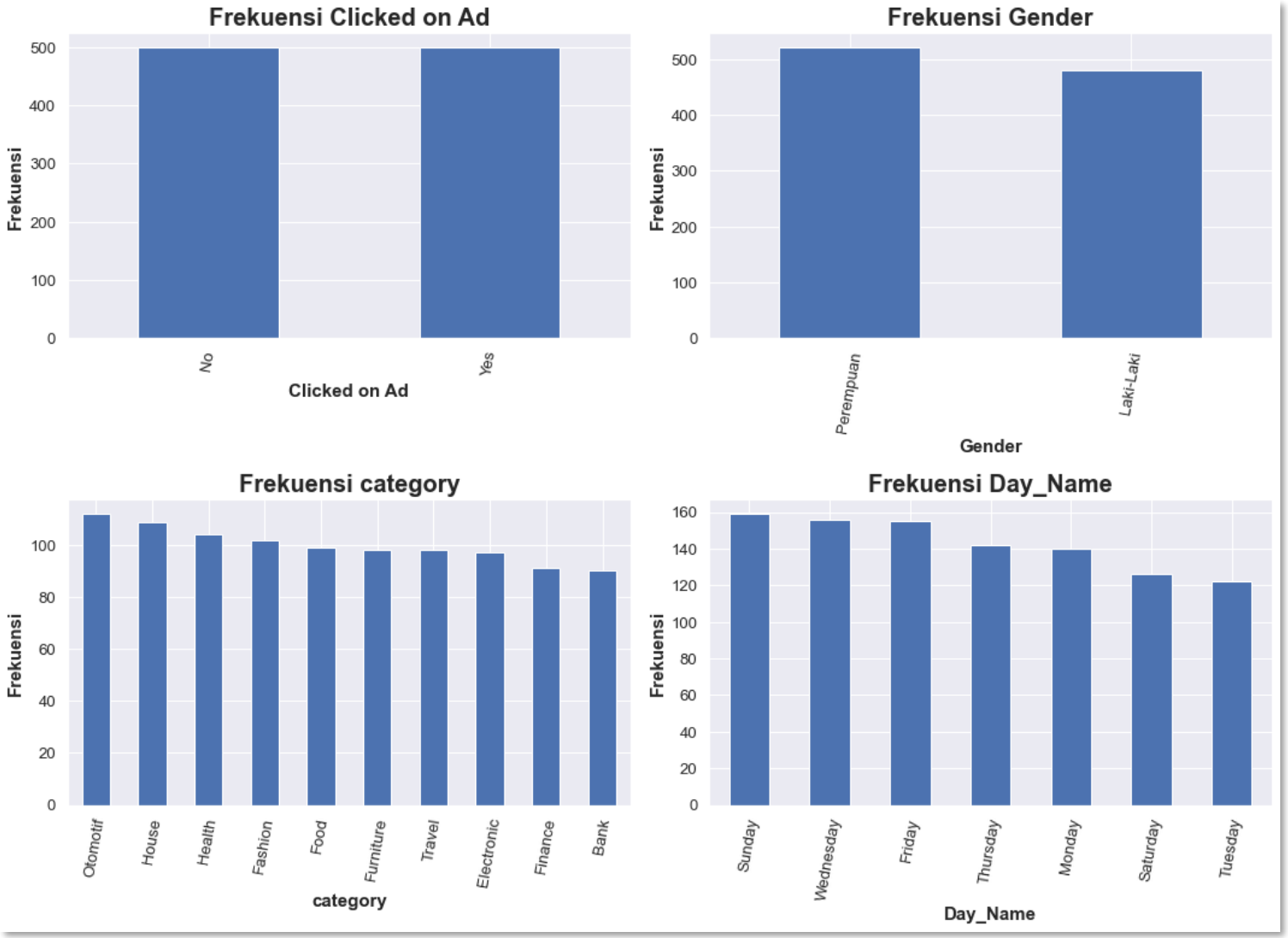
- 3 kolom/variable numerik terdapat missing value akan diinput dengan median
- Missing value pada variable Gender akan diisi dengan modusnya

- Tidak ada data duplicated
- Melakukan drop pada kolom/variable unamed

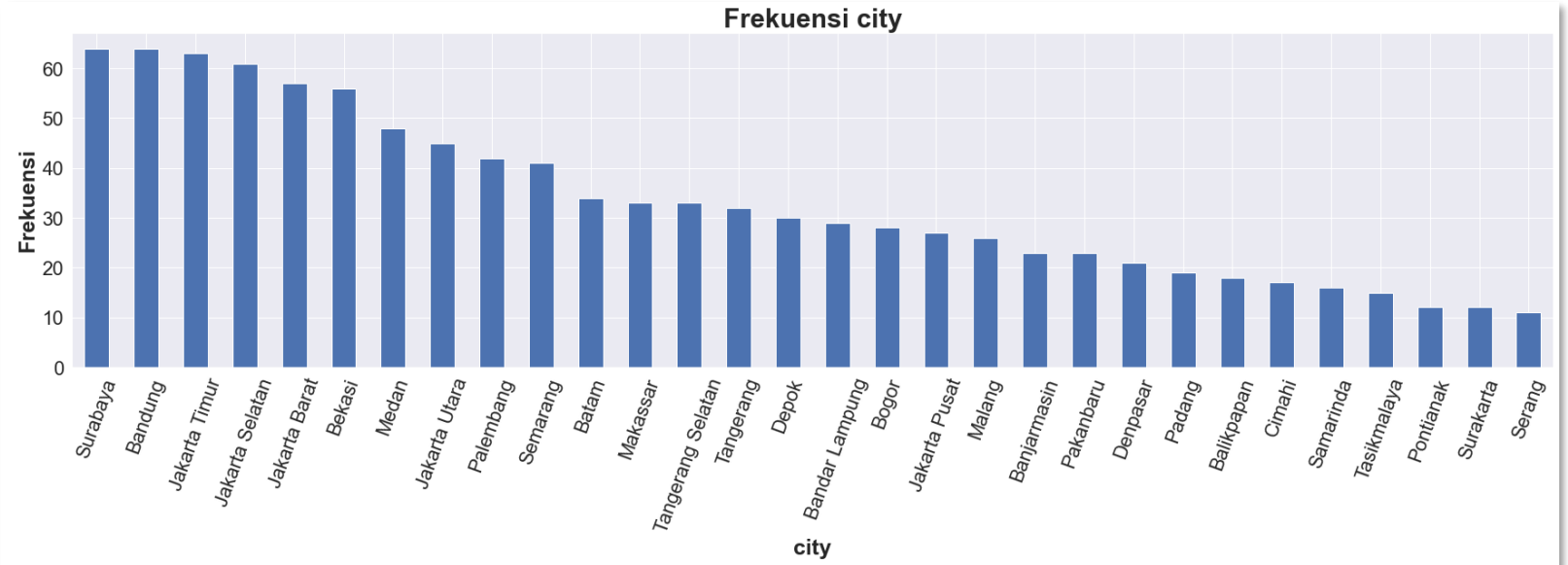
Univariate Analysis

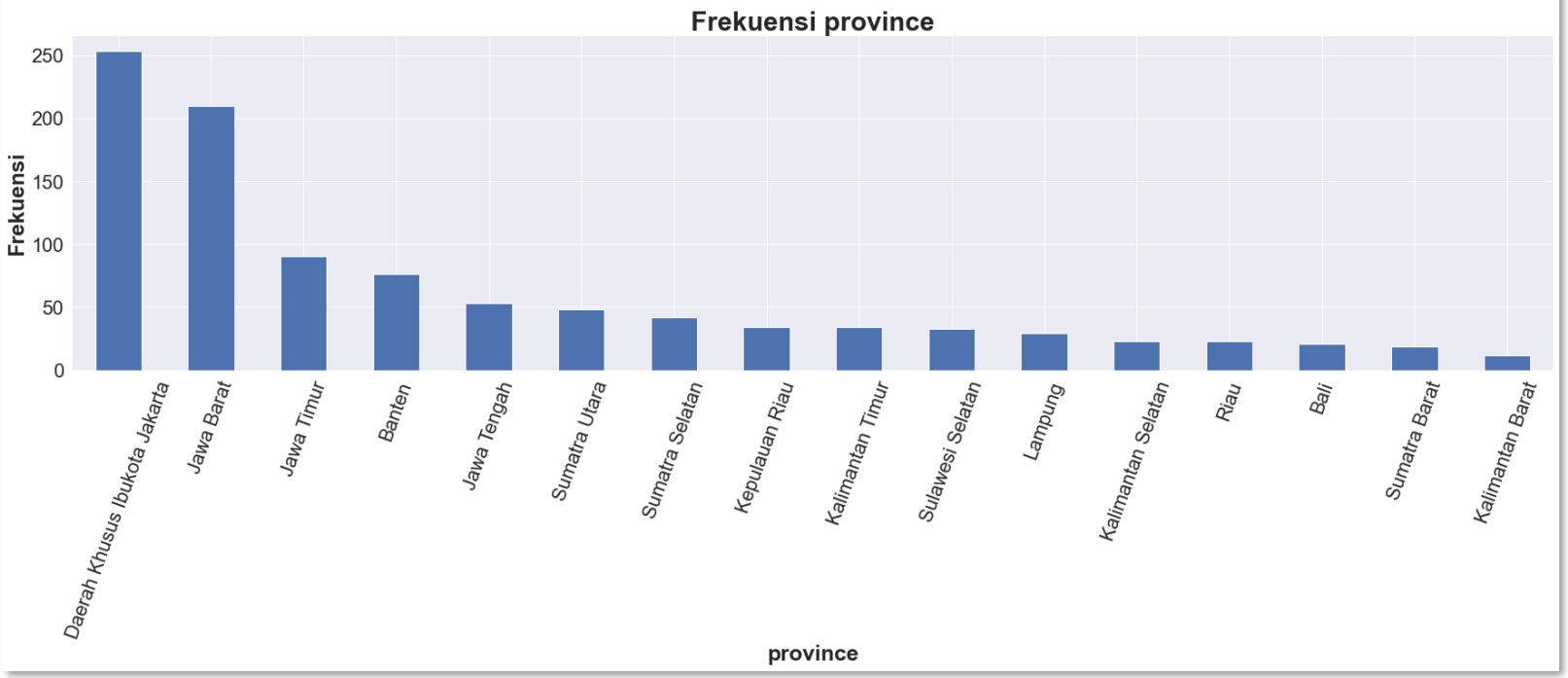


- Variabel Daily time spent on site dan Daily Internet Usage memiliki distribusi yang cenderung bimodal
- Variabel **Age** memiliki distribusi cenderung **positively skewed**, sedangkan variable **Area Income** memiliki distribusi **negatively skewed**
- Sebagian besar **customer** menghabiskan waktu berada di **platform/situs antara 60-80 menit**
- **Usia** customer adalah sekitar **20 - 60 tahun**
- Sebagian besar customer memiliki **pendapatan** antara **300 juta – 500 juta**
- **Penggunaan** harian **internet** adalah **100 – 250 menit** per hari.

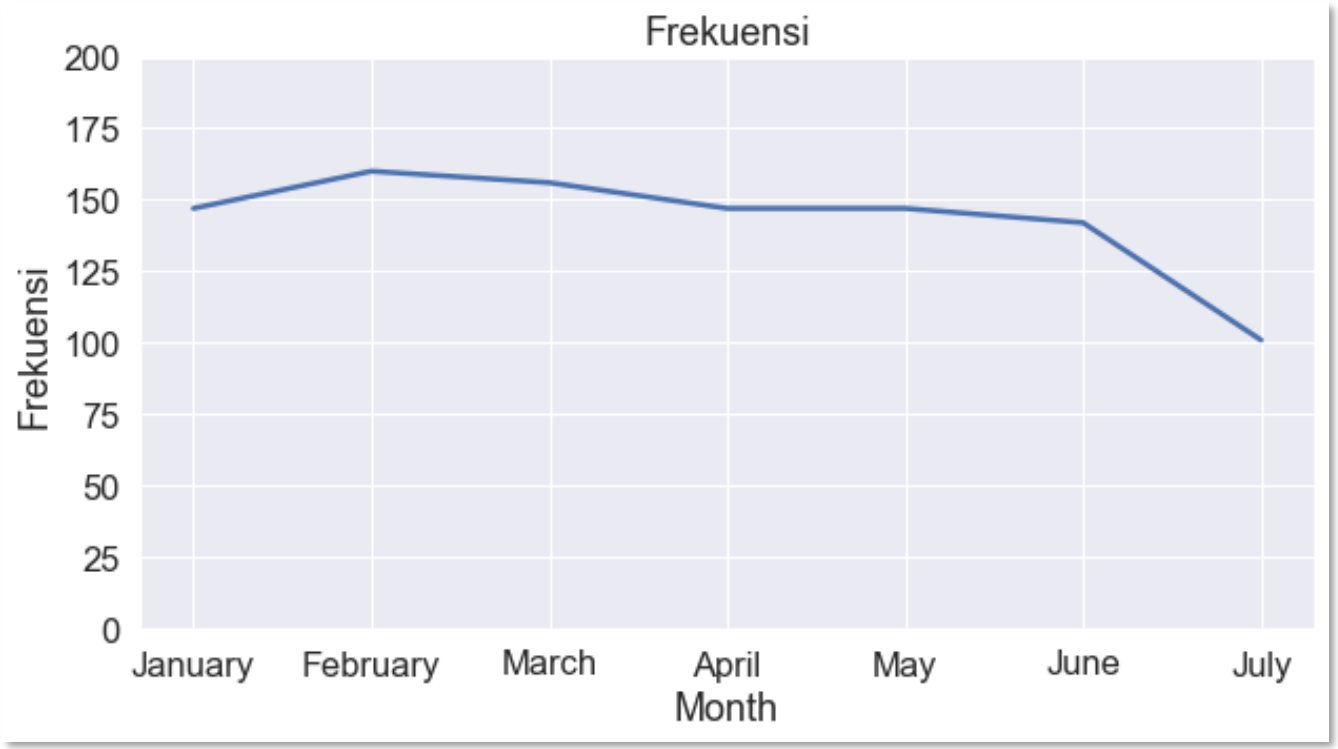


- Jumlah customer yang melakukan klik pada iklan adalah **sama** dengan yang tidak melakukan klik
- Frekuensi gender perempuan lebih banyak dibandingkan laki-laki, namun perbedaannya tidak terlalu signifikan
- **Otomotif, House dan Healt** adalah category iklan paling banyak
- Customer paling **banyak mengunjungi situs** pada hari **Minggu, Rabu dan Jumat**.



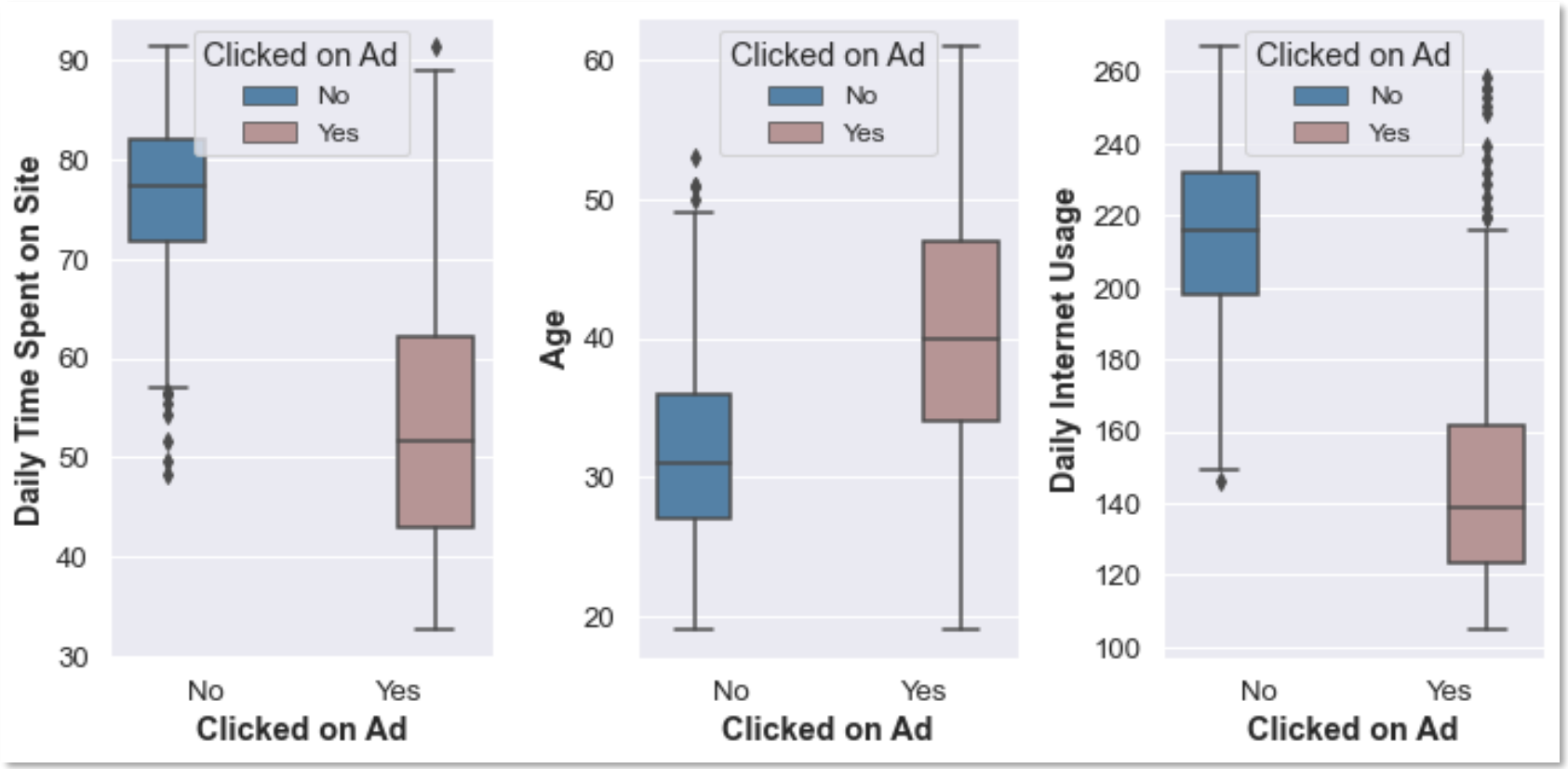


- Sebagian besar customer tinggal di kota Surabaya, Bandung dan Jakarta Timur
- Top 2 provinsi tempat tinggal customer adalah di DKI Jakarta dan Jawa Barat

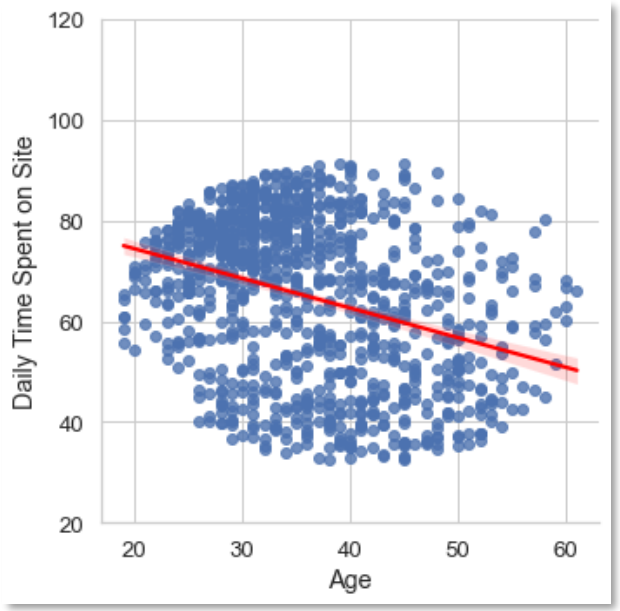


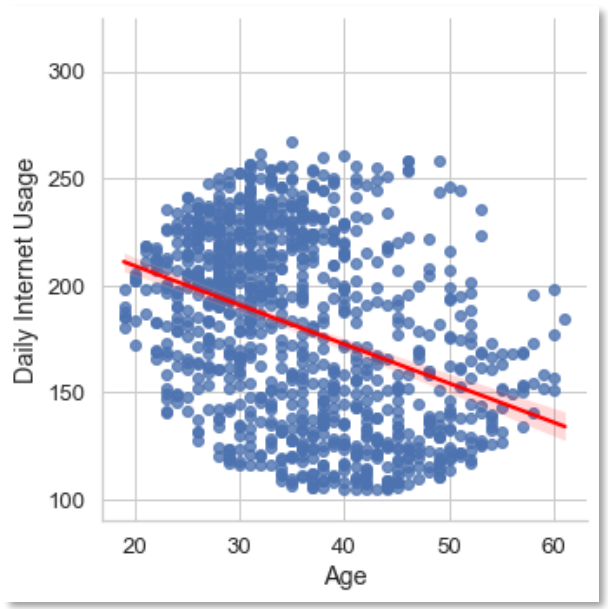
- Customer paling banyak mengunjungi situs/platform pada bulan Februari dan Maret, dan paling sedikit saat bulan Juli
- Tren yang terjadi dari bulan Januari hingga Juli cenderung terjadi penurunan

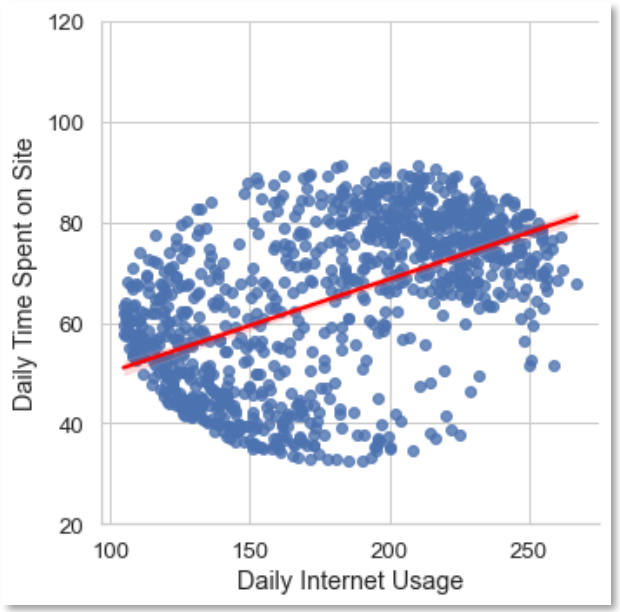
Bivariate Analysis



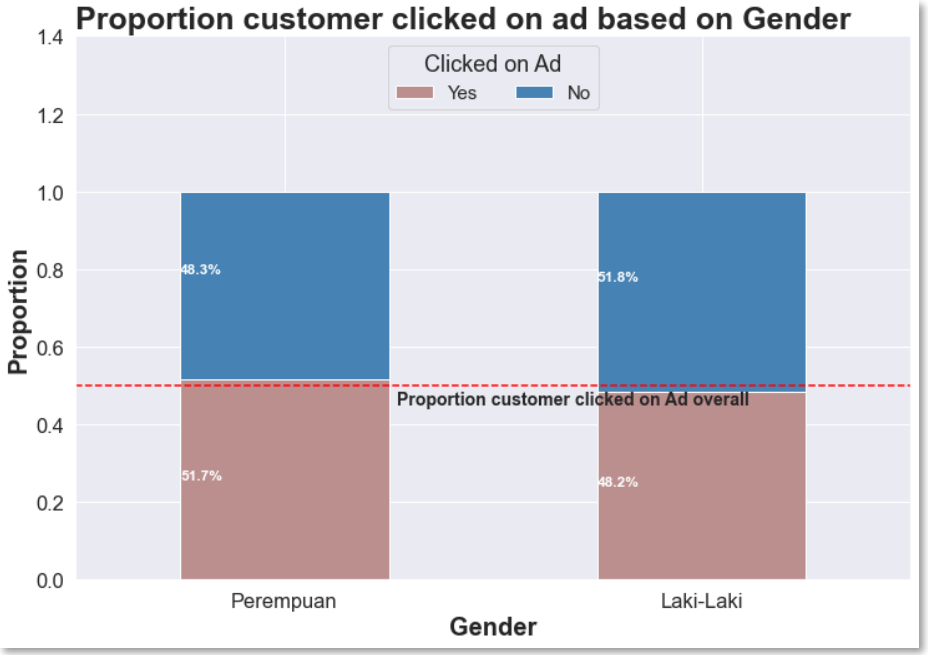
- Sebagian besar customer yang menghabiskan waktunya untuk berada di internet dan melakukan **clicked pada iklan** adalah customer yang **berusia di antara 35 - 50 tahun**. Sedangkan customer yang berusia antara 27-35 tahun sebagian besar tidak melakukan klik pada iklan.
- **Semakin lama** waktu yang dihabiskan untuk **berada di situs**, customer **cenderung tidak melakukan klik pada iklan**.
- Semakin **tinggi penggunaan internet,** cenderung **untuk tidak melakukan klik pada iklan** yang ada pada situs

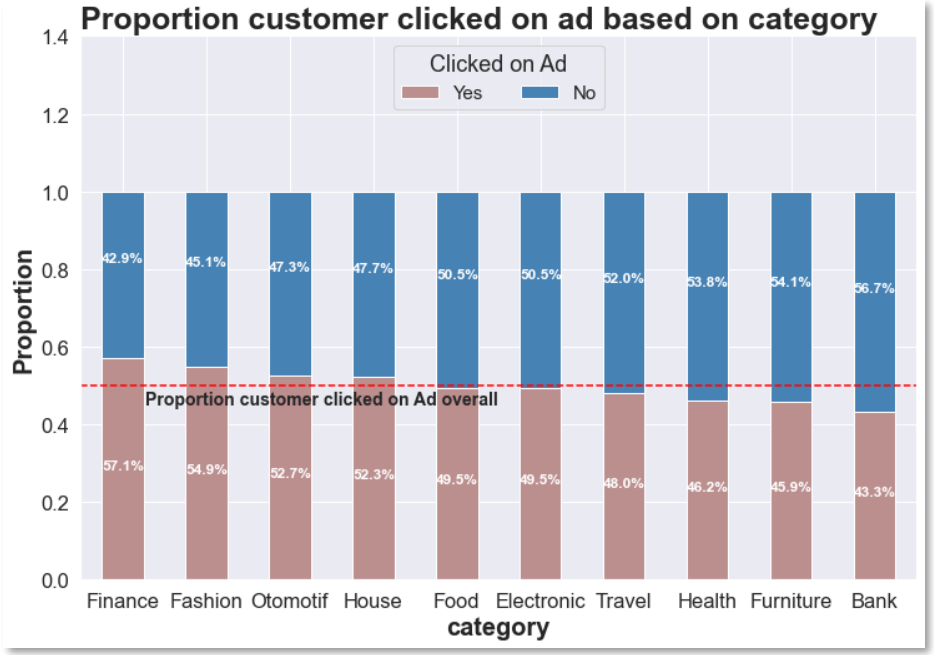


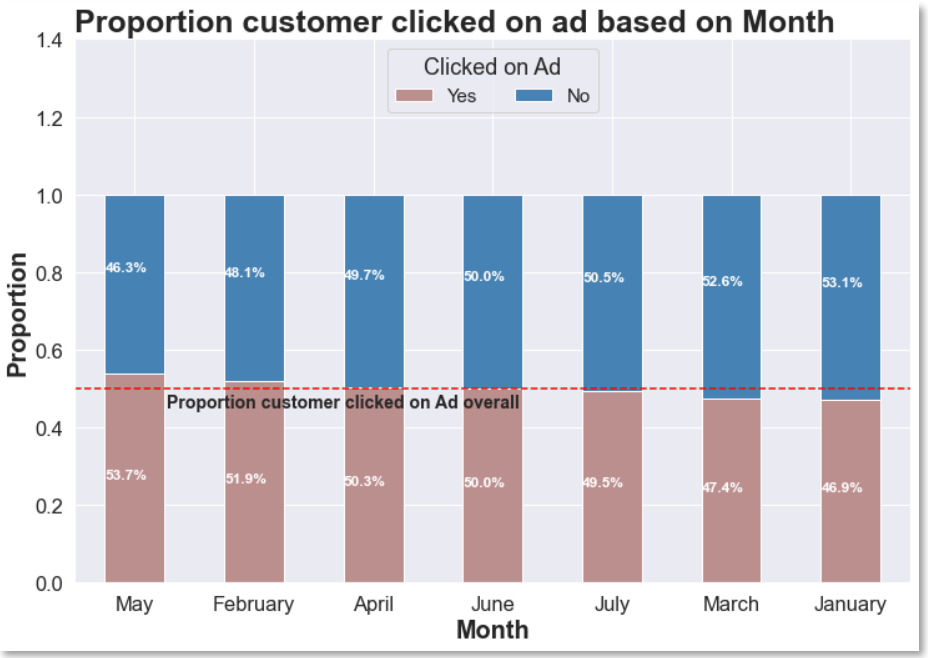


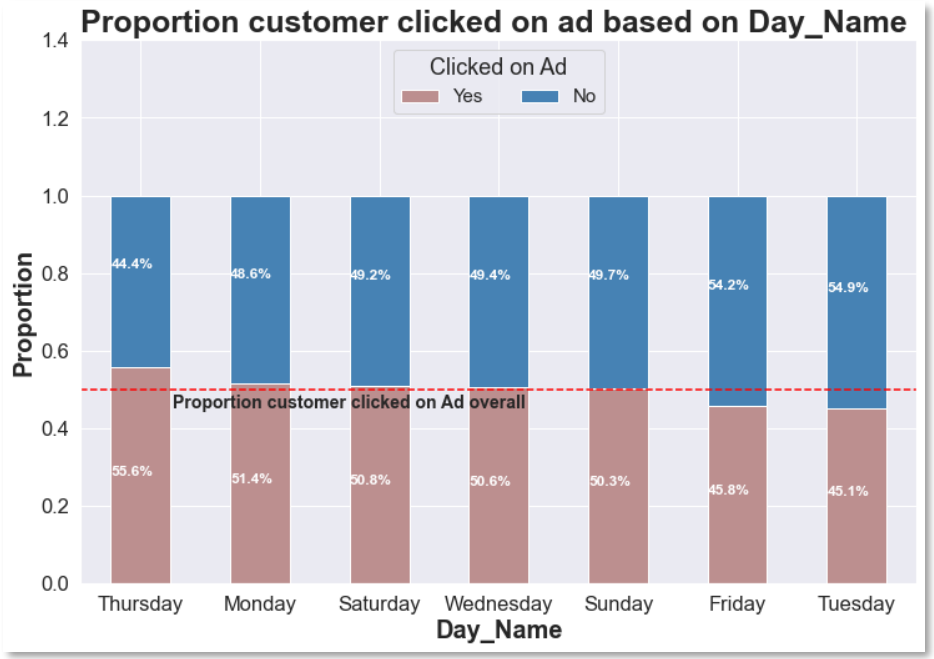


- Semakin tinggi usia customer, waktu harian yang dihabiskan untuk berada di situs juga semakin rendah
- Semakin tinggi usia customer penggunaan internet harian semakin rendah
- Semakin tinggi penggunaan harian internet, waktu harian yang dihabiskan untuk berada di situs juga semakin lama

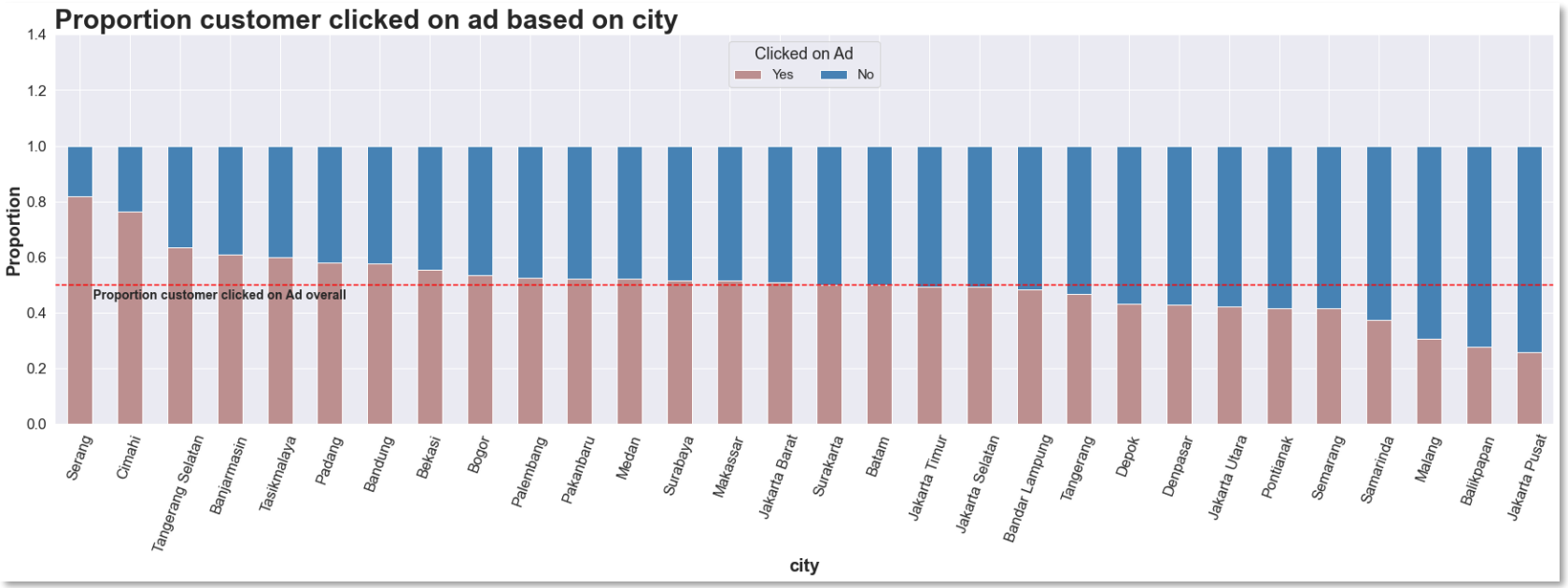


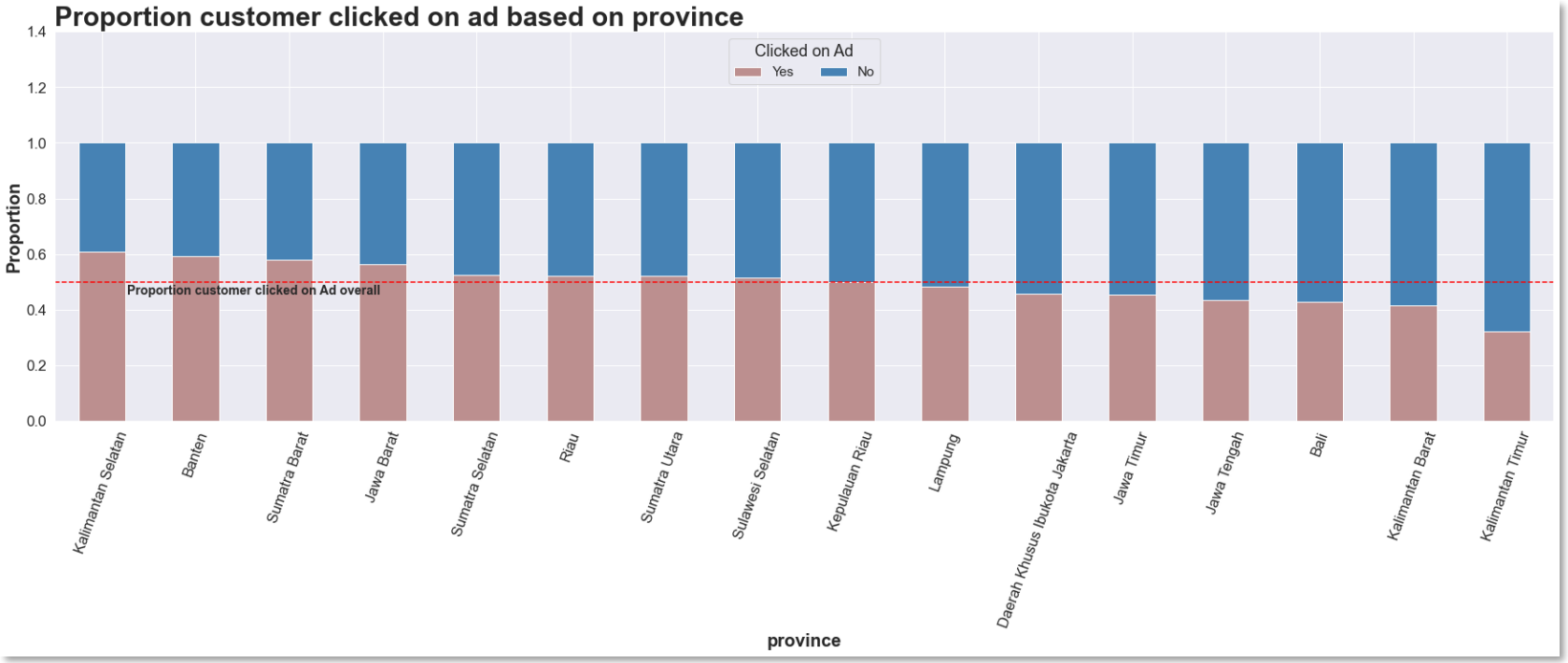






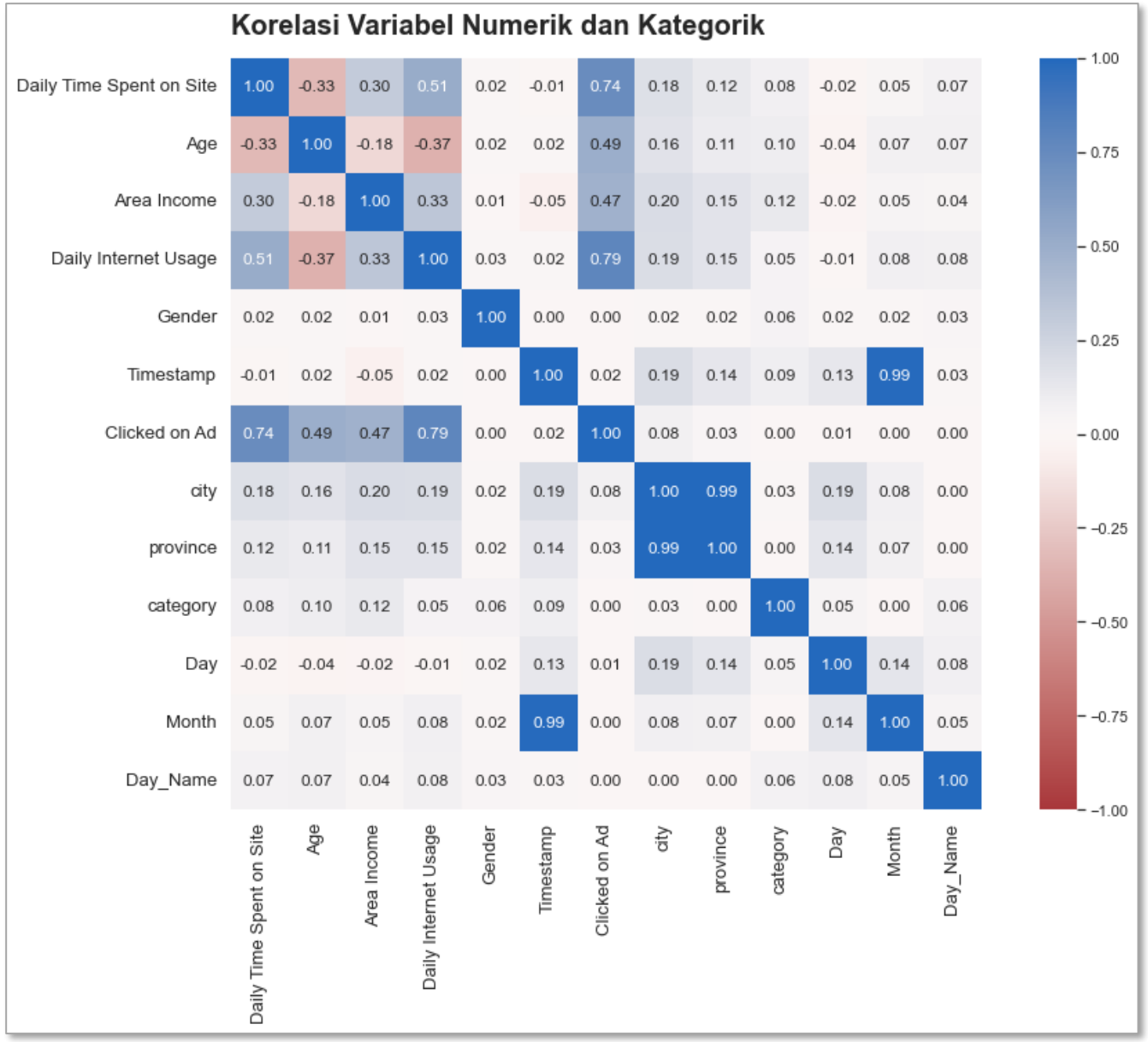
- Proporsi customer perempuan yang melakukan clicked pada Ad lebih tinggi dibandingkan laki-laki, namun perbedaannya tidak signifikan
- Proporsi customer melihat iklan pada **category Finance dan Fashion** adalah yang tertinggi pada category tersebut. Sedangkan pada category Bank adalah yang paling rendah
- Proporsi customer melihat iklan/ clicked on Ad pada bulan Januari adalah yang paling rendah dibandingkan dengan bulan yang lain
- Customer yang mengunjungi situs/platform pada **hari Kamis** secara **proporsi** lebih **tinggi** dibandingkan dengan hari lain.





- Proporsi customer melakukan clicked on Ad yang berasal dari kota Serang dan Cimahi adalah yang tertinggi. Sedangkan customer yang berasal dari Balikpapan dan Jakarta Pusat, proporsi untuk clicked on Ad adalah yang paling rendah
- Proporsi customer melakukan clicked on Ad yang berasal dari provinsi Kalimantan Selatan, Banten dan Sumatera Barat adalah yang tertinggi. Sedangkan proporsi customer untuk clicked on Ad yang berasal dari Kalimantan Timur adalah yang paling rendah.

Multivariate Analysis



- Ada beberapa variabel yang redundant antara lain : Month dengan Timestamp dan City dengan Province
- Variabel redundant yang memiliki korelasi paling lemah dengan variabel target (Clicked on Ad) akan didrop seperti variabel Month dan Province
- Variabel **Age dan income** memiliki **korelasi positive** yang sedang **dengan Clicked on Ad**
- Variabel **Daily time spent on site dan daily internet usage** memiliki **korelasi positive yang kuat** dengan **Clicked on Ad**
- Variabel daily time spent on site dan daily internet usage dengan age memiliki korelasi negative sedang
- Variabel daily time spent on site dan area income, memiliki korelasi positive sedang dengan variabel daily internet usage
- Variabel daily time spent on site dengan area income memiliki korelasi positive sedang

Data
Preprocessing

Link GitHub Repo

Data Cleansing & Preprocessing

Data Cleansing

- Cleaning data seperti handling missing value dan duplicated data telah dilakukan saat proses Exploratory Data Analysis

Ekstraksi Kolom/ Variabel Timestamp

- Ekstraksi kolom telah dilakukan saat proses EDA, dan setelah dilakukan uji statistik dan melihat korelasi fitur satu sama lain, didapatkan kolom/variable Day_Name yang dipertahankan

```
df['Timestamp’] = pd.to_datetime(df['Timestamp’],
                  errors='coerce')
df['Day'] = df['Timestamp'].dt.day
df['Month'] = df['Timestamp'].dt.month_name()
df['Day_Name'] = df['Timestamp'].dt.day_name()
```

Handling Outlier

- Tidak ada outlier yang dihapus karena bukan Global Outlier

Strategi feature encoding

- Day_Name dan Gender akan diencoding menggunakan Label Encoding
- City dan Category akan diencoding menggunakan One Hot Encoding
- Variabel numerik akan distandarisasi menggunakan Standard Scaler

Split Data

- Fitur yang digunakan adalah :

Day_Name, Gender, City, Category, Daily Time Spent on Site, Age, Area Income, Daily Internet Usage

- Variabel Target adalah Clicked on Ad
- Data dipisah menjadi train dan test, masing-masing 80% dan 20% dari total data

```
X_train, X_test, y_train, y_test =
train_test_split(X,y,stratify=y,test_size=0.2,
random_state=42)
```

Link GitHub Repo

Modeling

Link GitHub Repo

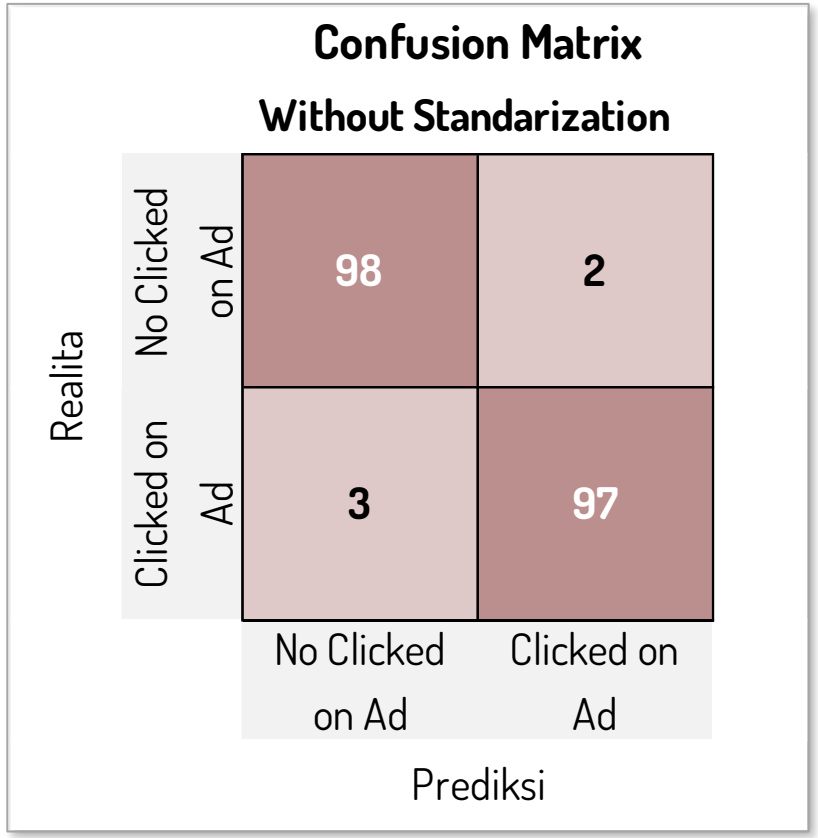
Hasil Experiment

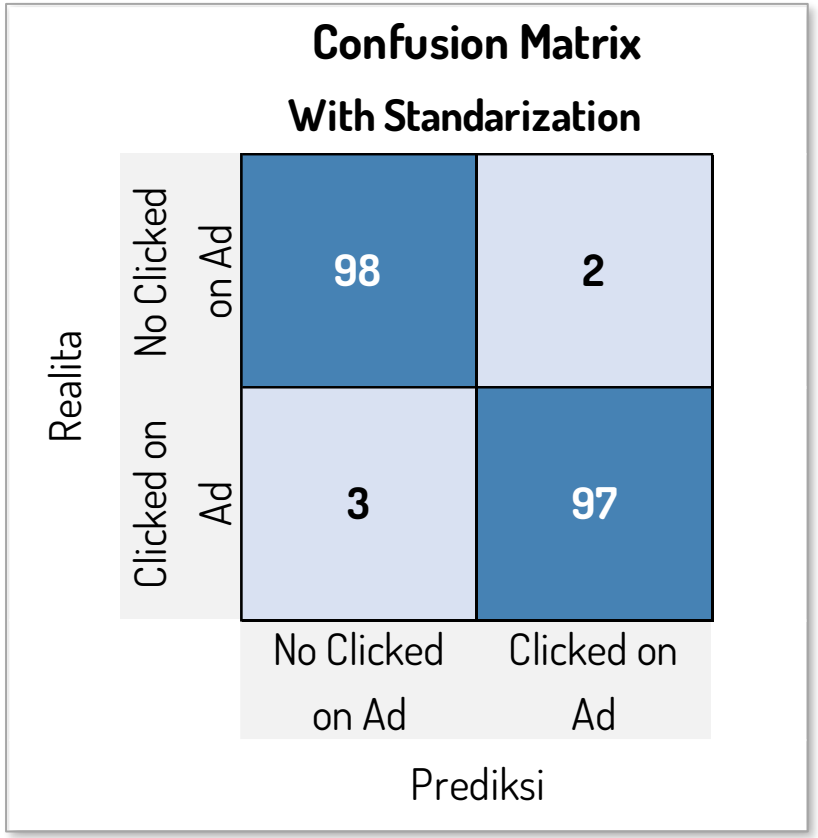
Metric yang digunakan adalah F1 Score

$$F1 - Score = \frac{2\,TP}{2\,TP + FN + FP}$$

Membuat model yang sekecil mungkin salah prediksi pada customer yang kemungkinan clicked on Ad serta salah prediksi pada customer yang tidak melakukan clicked on Ad. Sehingga tidak kehilangan potensial customer dan meminimumkan customer yang tidak melakukan clicked on Ad.

Dengan menggunakan algoritma Adaboost dan telah dilakukan tuning hyperparameter, tidak ada perbedaan hasil dari confusion matrix dari kedua experiment, score di beberapa metrics pun juga tidak ada perbedaan.



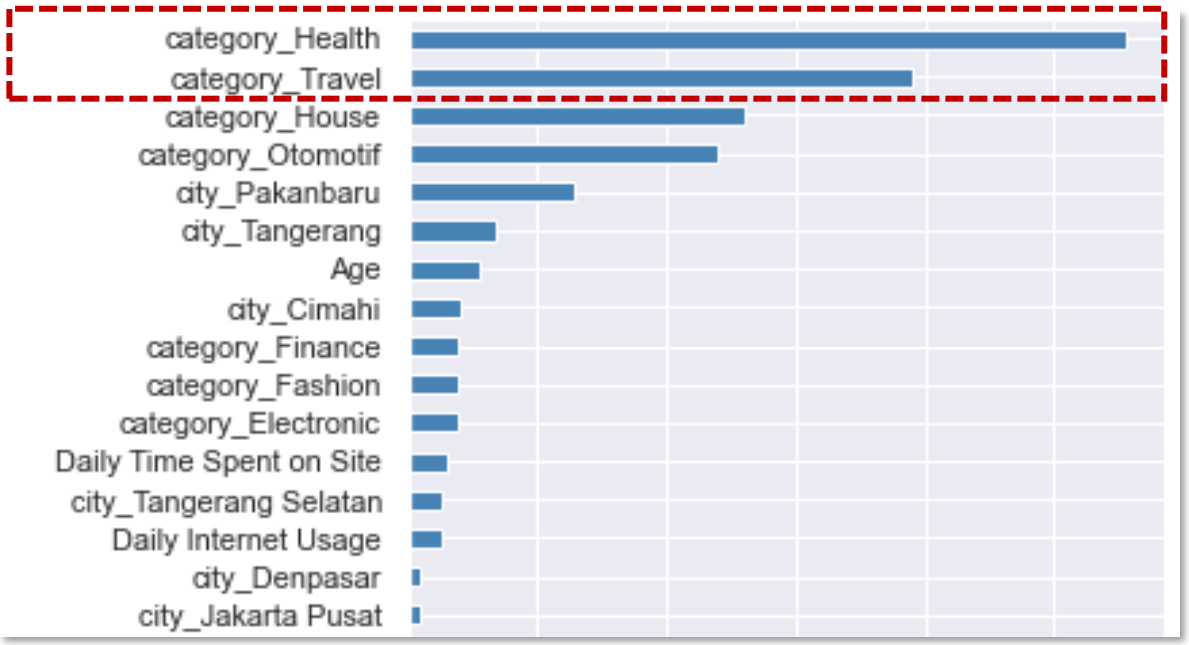


Model Adaboost in Data Test	Accuracy	F1	ROC_AUC
With Standarizationn	0.975	0.975	0.985
Without Standarization	0.975	0.975	0.985

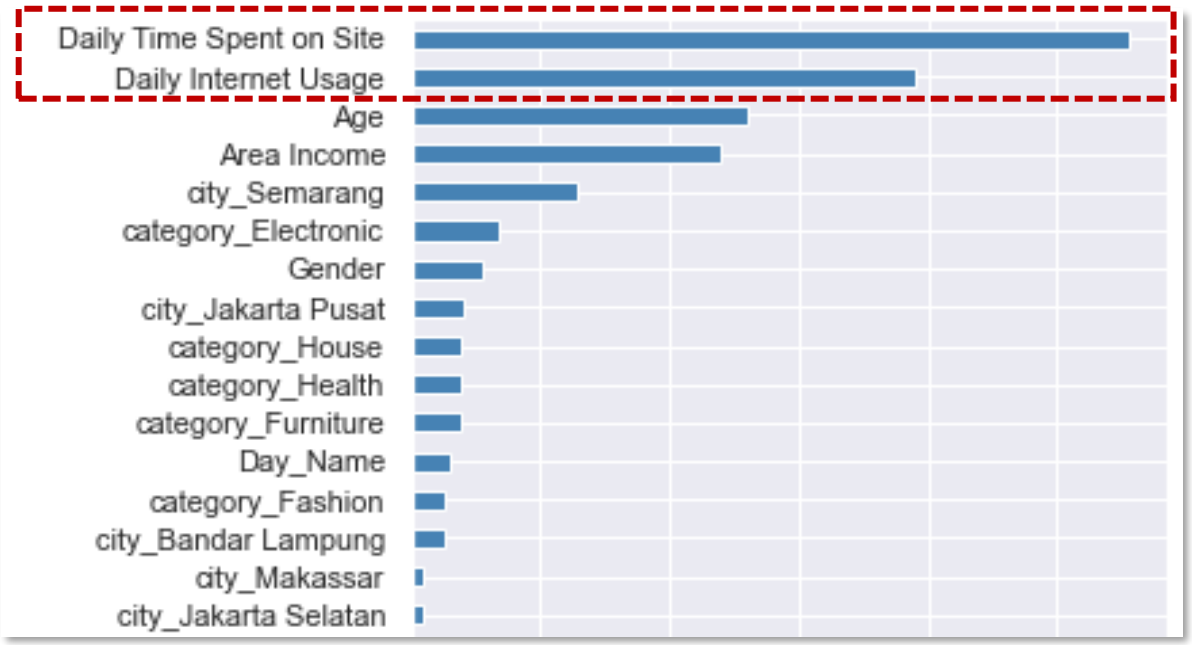
Link GitHub Repo

Feature Importance

Tanpa Standarisasi



dengan Standarisasi



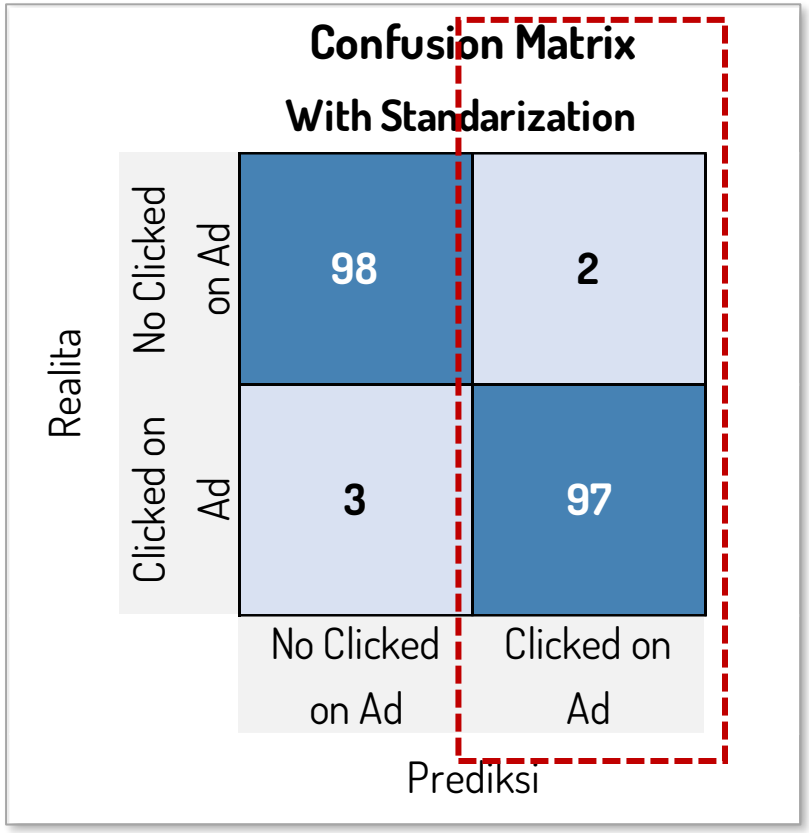
Ada perbedaan feature importance diantara keduanya. Jika tanpa standarisasi 2 fitur utama adalah category Health dan Travel, sedangkan jika menggunakan standarisasi 2 fitur utama keberhasilan marketing adalah **penggunaan harian internet dan waktu yang dihabiskan saat berada di situs/platform**

Link GitHub Repo

Business Simulation
& Recommendation

Link GitHub Repo

Business Simulation



Asumsi : Revenue dan Marketing Cost. Customer yang clicked on Ad akan melakukan transaksi

Revenue per customer : Rp 20,000

Marketing cost per customer : Rp 15,000

Keterangan	Tanpa Machine Learning	dengan Machine Learning
Jumlah Customer	200	200
Jumlah Customer clicked Ads	100	99
Conversion Rate	50%	98%
Total Profit	Rp 2,000,000	Rp 1,980,000
Total Marketing cost	Rp 3,000,000	Rp 1,485,000
Profit	**Rp -1,000,000**	**Rp 495,000**

Source : smallbusiness.chron

Kemungkinan customer melakukan clicked on Ad dan bertransaksi adalah 50:50 jika tidak menggunakan model (berdasarkan data yang ada 50% customer akan clicked on Ad). Jika menggunakan model pada Machine Learning, dari 200 customer diprediksi yang akan melakukan clicked on Ad adalah 99 orang, sehingga bisa menghemat cost marketing hingga setengahnya jika tanpa model.

Link GitHub Repo

Business Recommendation

Customer kategori A merupakan customer yang berusia muda yang memiliki history penggunaan internet harian 150 menit ke atas dan menghabiskan waktu berada di situs selama > 60 menit

Customer kategori B customer yang berusia lebih tua, memiliki history penggunaan internet harian < 175 menit dan menghabiskan waktu berada di situs/platform selama < 60 menit.

Customer berusia muda dan memiliki history penggunaan internet lebih lama cenderung tidak melakukan klik/melihat iklan. Mungkin saja hal tersebut terjadi karena terbiasa melihat iklan di internet sehingga mengabaikannya, berbeda halnya dengan customer kategori B (usia lebih tua dari customer kategori A) memiliki history penggunaan internet lebih rendah namun melakukan klik pada iklan.

1. Customer kategori B merupakan customer yang potensial untuk menerima marketing campaign melalui clicked on Ad. Sehingga dapat memasang iklan digital pada situs/platform yang sering dikunjungi customer kategori tersebut.
2. Mengoptimalkan iklan pada kategori Finance, fashion dan otomotif. Karena customer sering melakukan klik pada iklan kategori tersebut
3. Iklan yang ditampilkan pada hari Kamis lebih sering diklik/dilihat customer sehingga dapat mengoptimalkan pada hari tersebut, karena hari Kamis dirasa cukup potensial untuk melakukan marketing campaign melalui iklan digital.

Link GitHub Repo

Terima kasih