Predict Customer Clicked Ads Classification by using Machine Learning

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

Created by:
Muhammad Miftah Thaha
muhammadmiftaht@gmail.com
https://www.linkedin.com/in/miftahthaha/
https://github.com/miftahthaha

Miftah is a recent graduate with a Bachelor's degree Electrical Engineer who have interest in Data Analytics and Science and have a strong foundation in statistical modeling, data analysis, and programming. As a Junior Data Scientist, he has experience through his final project in building and implementing machine learning models, analyzing complex data sets, and creating visualizations to communicate insights. He is a fast learner with excellent problem-solving skills and a passion for using data to drive business decisions. In addition, he possess strong communication and collaboration skills, having worked on multiple team projects during his studies. With a drive to excel in his field, Miftah is seeking an opportunity to contribute his skills and knowledge to a dynamic and innovative organization as a Junior Data Scientist.

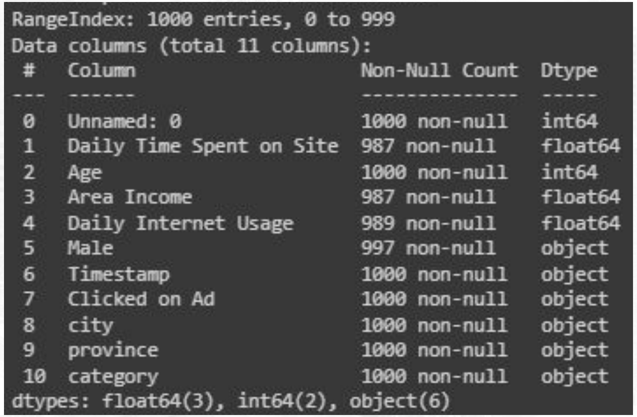
Overview

“A company in Indonesia wants to determine the effectiveness of an advertisement they are running. This is important for the company to understand how successful their marketed advertisement is in reaching customers and attracting them to view the advertisement.

By analyzing historical advertisement data and identifying insights and patterns, it can help the company determine their marketing targets. The focus of this case is to create a machine learning classification model that can determine the right target customers.”

Customer Type and Behaviour Analysis on Advertisement

Dataset



```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Unnamed: 0                1000 non-null   int64
 1   Daily Time Spent on Site  987 non-null    float64
 2   Age                       1000 non-null   int64
 3   Area Income               987 non-null    float64
 4   Daily Internet Usage      989 non-null    float64
 5   Male                      997 non-null    object
 6   Timestamp                 1000 non-null   object
 7   Clicked on Ad             1000 non-null   object
 8   city                      1000 non-null   object
 9   province                  1000 non-null   object
 10  category                  1000 non-null   object
dtypes: float64(3), int64(2), object(6)
```

- Description

  Dataset that contains information related to personal browsing history made by Ads Company.

- Shape

  1000 Row and 11 Columns (Feature)

- Datatypes

  Float64 (3 Feature), Int64 (2 Feature), object (6 Feature)

- Missing Values

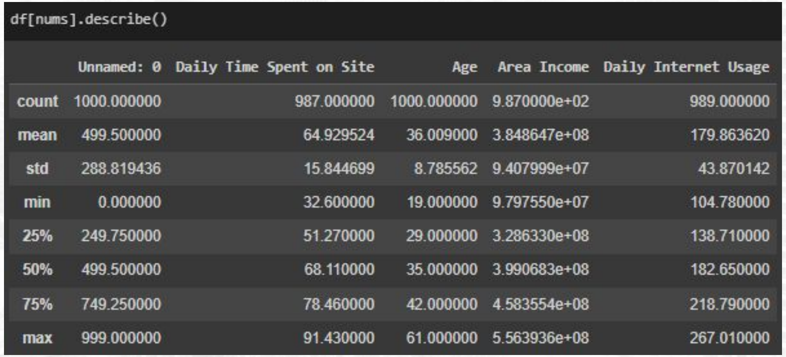
  Detected in 6 Column

For more details, you can see all file here and code here

Exploration Data Analysis (EDA)

Descriptive Statistic

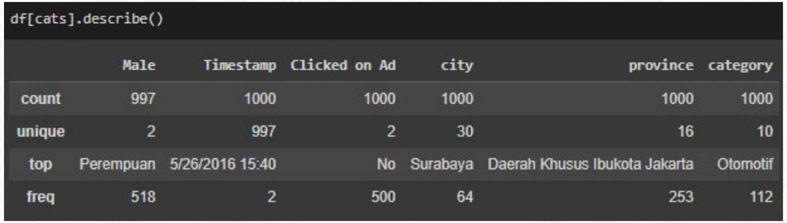
- Data Numeric



```
df[nums].describe()
```

```
        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
```

- Data Categoric



```
df[cats].describe()
```

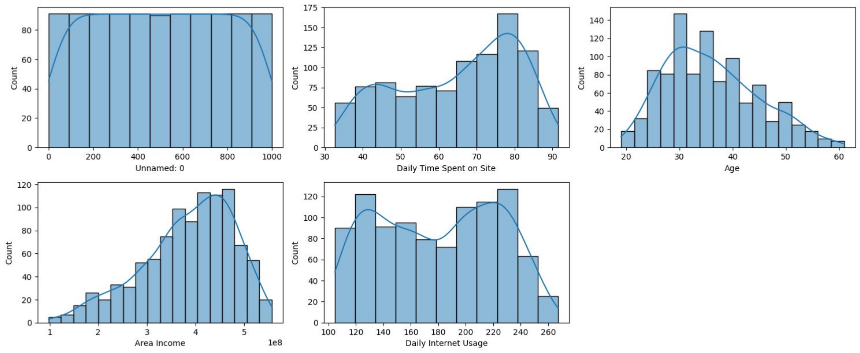
```
             Male        Timestamp  Clicked on Ad      city                    province  category
count         997             1000           1000      1000                        1000      1000
unique          2              997              2        30                          16        10
top     Perempuan  5/26/2016 15:40             No  Surabaya  Daerah Khusus Ibukota Jakarta  Otomotif
freq          518                2            500        64                         253       112
```

For more details, you can see all file here and code here

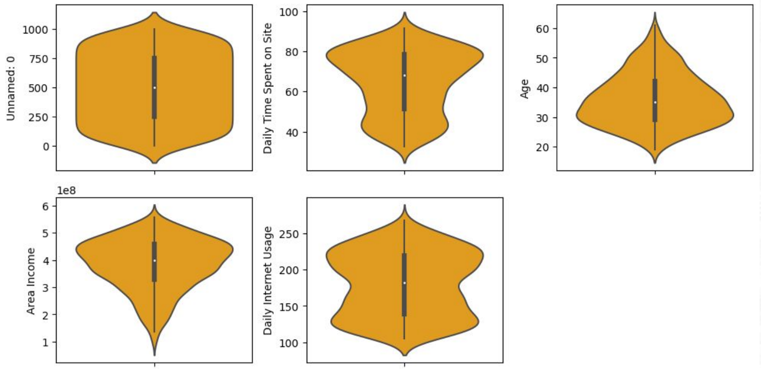
Customer Type and Behaviour Analysis on Advertisement

Univariate Analysis

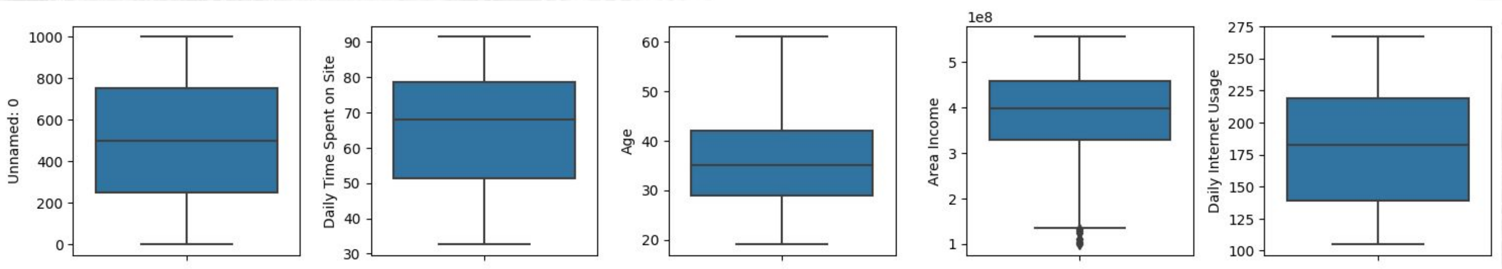
- Histogram



- Violin Plot



- Boxplot



For more details, you can see all file here and code here

Univariate Analysis

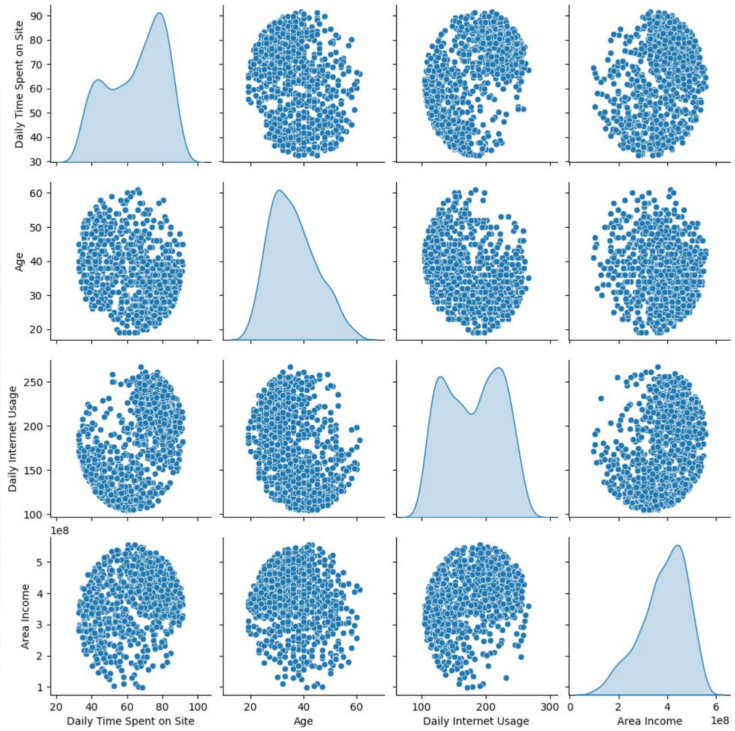
By identifying these key findings through univariate analysis, we can gain a better understanding of the dataset and use this information to inform future analyses and decision-making.

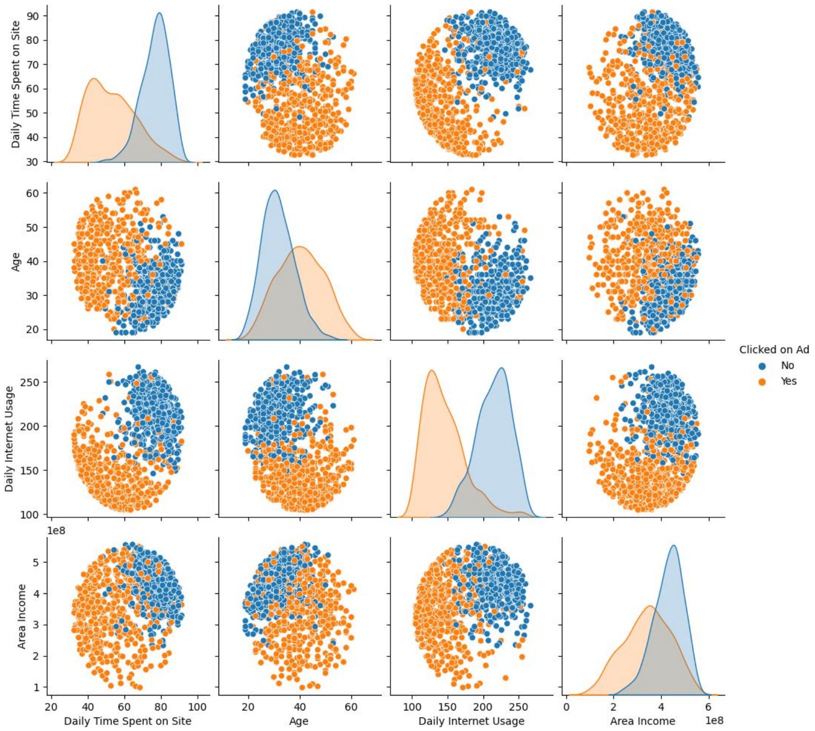
- Outliers: Outliers were only found in the "Area Income" feature, indicating that this feature may have extreme values that are significantly different from the other data points in the dataset.
- Skewed distribution: The distribution of "Age" and "Area Income" features were left and slightly skewed respectively, suggesting that these features may not follow a normal distribution.
- Bimodal distribution: The distributions of "Daily Time Spent on Site" and "Daily Internet Usage" features showed a bimodal distribution, indicating that these features may have two distinct groups of data points or behaviors.

For more details, you can see all file here and code here

Customer Type and Behaviour Analysis on Advertisement

Bivariate Analysis





For more details, you can see all file here and code here

Customer Type and Behaviour Analysis on Advertisement

Bivariate Analysis

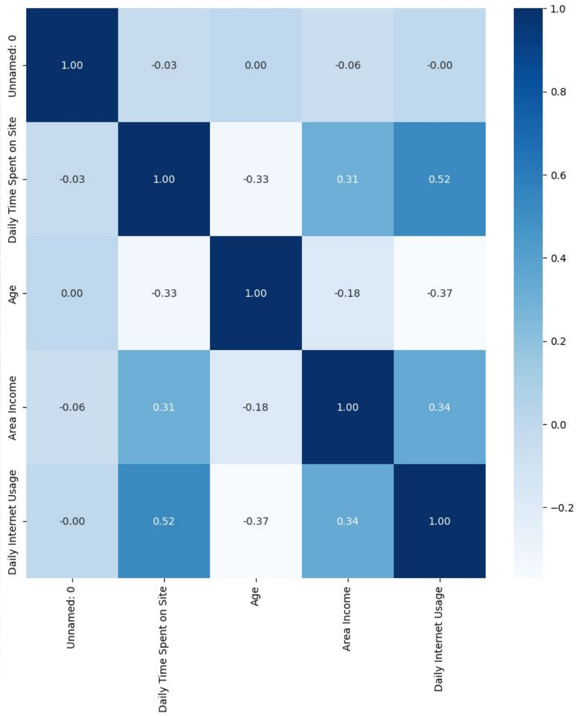
By identifying these key findings through bivariate analysis, we can gain a better understanding of how different variables relate to each other and how they impact customer behavior. This information can help us optimize our marketing strategies to reach our target audience more effectively.

- Daily Time Spent on Site: Customers who spend between 35-45 minutes on the site are more likely to click on the ad. However, those who spend 70-80 minutes on the site are less likely to click on the ad.
- Age: Customers between the ages of 35-45 are more likely to click on the ad, while those between 25-35 are less likely to click on the ad.
- Daily Internet Usage: Customers who spend between 100-150 minutes on the internet daily are more likely to click on the ad. However, those who spend between 175-225 minutes on the internet daily are less likely to click on the ad.
- Area Income: Customers with an income range of around 380-460 million are less likely to click on the ad compared to those with other income ranges.

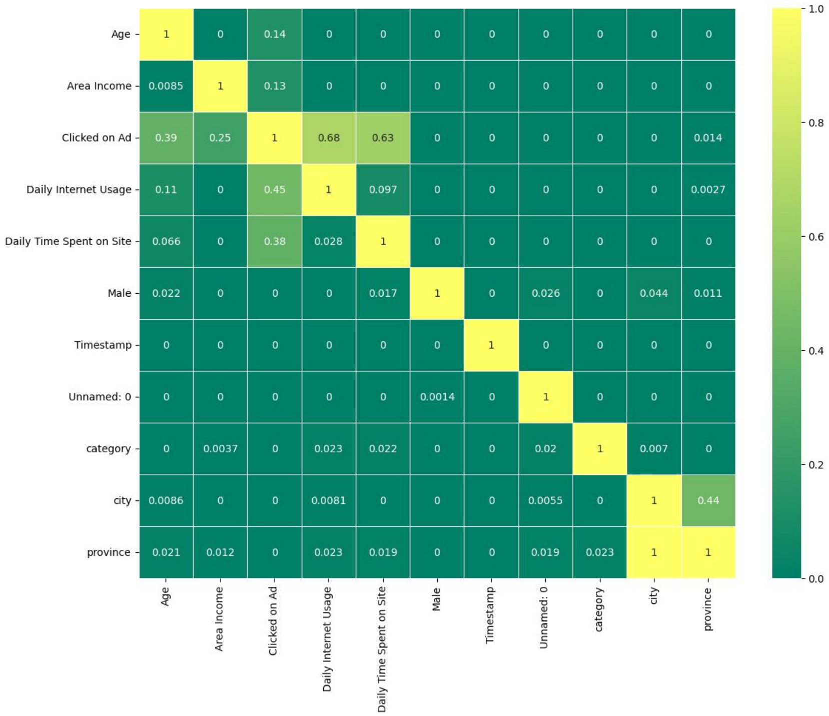
For more details, you can see all file here and code here

Customer Type and Behaviour Analysis on Advertisement

Multivariate Analysis



	Unnamed: 0	Daily Time Spent on Site	Age	Area Income	Daily Internet Usage
Unnamed: 0	1.00	-0.03	0.00	-0.06	-0.00
Daily Time Spent on Site	-0.03	1.00	-0.33	0.31	0.52
Age	0.00	-0.33	1.00	-0.18	-0.37
Area Income	-0.06	0.31	-0.18	1.00	0.34
Daily Internet Usage	-0.00	0.52	-0.37	0.34	1.00



	Age	Area Income	Clicked on Ad	Daily Internet Usage	Daily Time Spent on Site	Male	Timestamp	Unnamed: 0	category	city	province
Age	1	0	0.14	0	0	0	0	0	0	0	0
Area Income	0.0085	1	0.13	0	0	0	0	0	0	0	0
Clicked on Ad	0.39	0.25	1	0.68	0.63	0	0	0	0	0	0.014
Daily Internet Usage	0.11	0	0.45	1	0.097	0	0	0	0	0	0.0027
Daily Time Spent on Site	0.066	0	0.38	0.028	1	0	0	0	0	0	0
Male	0.022	0	0	0	0.017	1	0	0.026	0	0.044	0.011
Timestamp	0	0	0	0	0	0	1	0	0	0	0
Unnamed: 0	0	0	0	0	0	0.0014	0	1	0	0	0
category	0	0.0037	0	0.023	0.022	0	0	0.02	1	0.007	0
city	0.0086	0	0	0.0081	0	0	0	0.0055	0	1	0.44
province	0.021	0.012	0	0.023	0.019	0	0	0.019	0.023	1	1

For more details, you can see all file here and code here

Multivariate Analysis

By analyzing the correlations between these variables, we can gain insights into how different factors impact customer behavior and which factors may be more influential in predicting ad clicks. This information can be used to optimize marketing strategies and improve overall campaign performance.

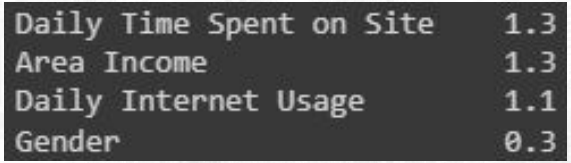
- Column Daily Internet Usage has a positive correlation with Daily Time Spent on Site equal to 0.52 and 0.34 with column Area Income. Additionally, Daily Internet Usage is negatively correlated with Age with a correlation of -0.37.
- Column Daily Time Spent on Site has a negative correlation with Age equal to -0.33, and has a positive correlation with Daily Internet Usage but a weak positive correlation with Area Income.
- Column Area Income is weakly positively correlated with Daily Time Spent on Site and Daily Internet Usage with correlation coefficients of 0.24 and 0.34, respectively.
- Feature Age has a weak negative correlation with all three features: Daily Time Spent on Site, Area Income, and Daily Internet Usage.
- The pairplot above reveals that customers who clicked on the ad and those who did not can be grouped clearly for columns Daily Time Spent on Site and Daily Internet Usage.

For more details, you can see all file here and code here

Data Cleaning & Preprocessing

- Handling Null Value

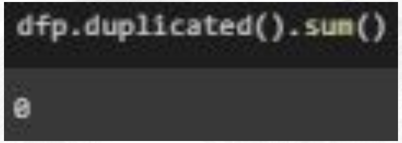
4 columns had null values on Daily Time Spent on Site, Area Income, Daily Internet Usage, and Gender Column. Impute numerical and categorical columns with median and mode values respectively.



```
Daily Time Spent on Site    1.3
Area Income                 1.3
Daily Internet Usage        1.1
Gender                      0.3
```

- Handling Duplicated Value

No Duplicated Values

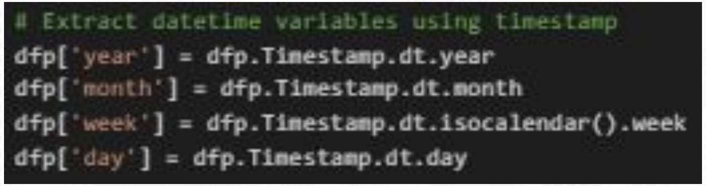


```
dfp.duplicated().sum()

0
```

- Feature Extraction

Extracted new columns from Timestamp Column.



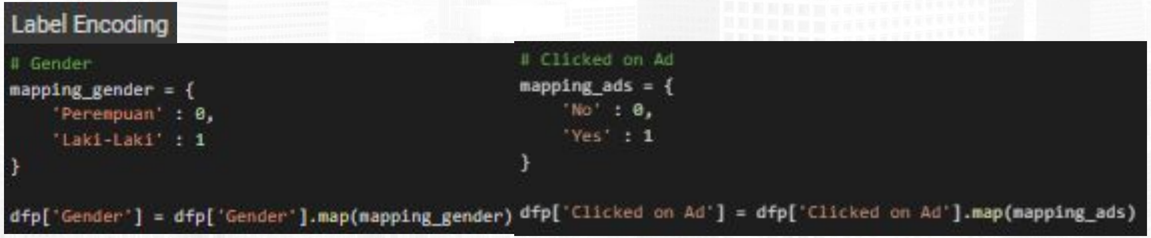
```
# Extract datetime variables using timestamp
dfp['year'] = dfp.Timestamp.dt.year
dfp['month'] = dfp.Timestamp.dt.month
dfp['week'] = dfp.Timestamp.dt.isocalendar().week
dfp['day'] = dfp.Timestamp.dt.day
```

For more details, you can see all file here and code here

Data Cleaning & Preprocessing

- Feature Encoding

Label encoding will be applied to the Gender and Clicked to Ads column, also One Hot encoding will be applied to city and province column as it will be used in the modeling process.



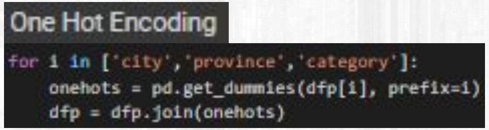
Label Encoding

```
# Gender
mapping_gender = {
    'Perempuan' : 0,
    'Laki-Laki' : 1
}

dfp['Gender'] = dfp['Gender'].map(mapping_gender)
```

```
# Clicked on Ad
mapping_ads = {
    'No' : 0,
    'Yes' : 1
}

dfp['Clicked on Ad'] = dfp['Clicked on Ad'].map(mapping_ads)
```

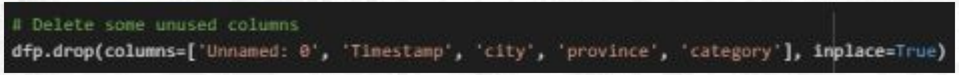


One Hot Encoding

```
for i in ['city','province','category']:
    onehots = pd.get_dummies(dfp[i], prefix=i)
    dfp = dfp.join(onehots)
```

- Feature Selection

Drop Unused columns.

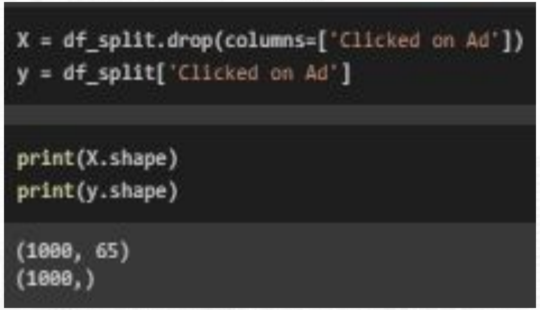


```
# Delete some unused columns
dfp.drop(columns=['Unnamed: 0', 'Timestamp', 'city', 'province', 'category'], inplace=True)
```

For more details, you can see all file here and code here

Data Cleaning & Preprocessing

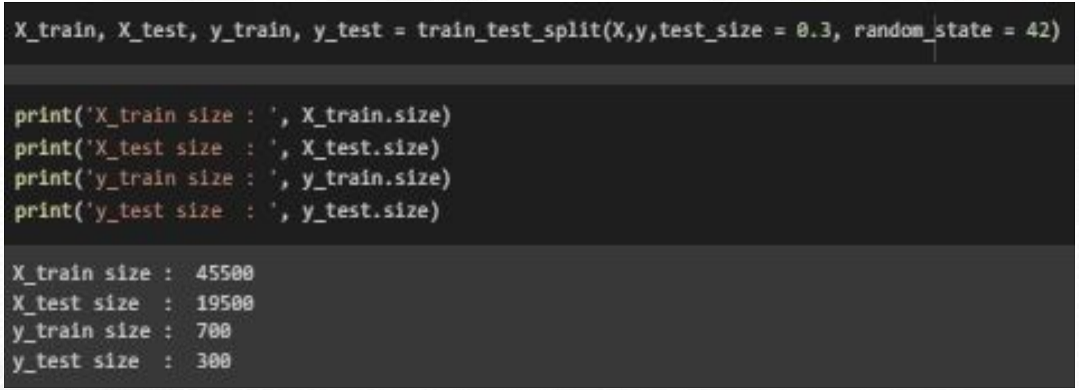
- Split Data



```
X = df_split.drop(columns=['Clicked on Ad'])
y = df_split['Clicked on Ad']
```

```
print(X.shape)
print(y.shape)
```

```
(1000, 65)
(1000,)
```



```
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.3, random_state = 42)
```

```
print('X_train size : ', X_train.size)
print('X_test size  : ', X_test.size)
print('y_train size : ', y_train.size)
print('y_test size  : ', y_test.size)
```

```
X_train size :  45500
X_test size  :  19500
y_train size :  700
y_test size  :  300
```

For more details, you can see all file here and code here

Data Modeling

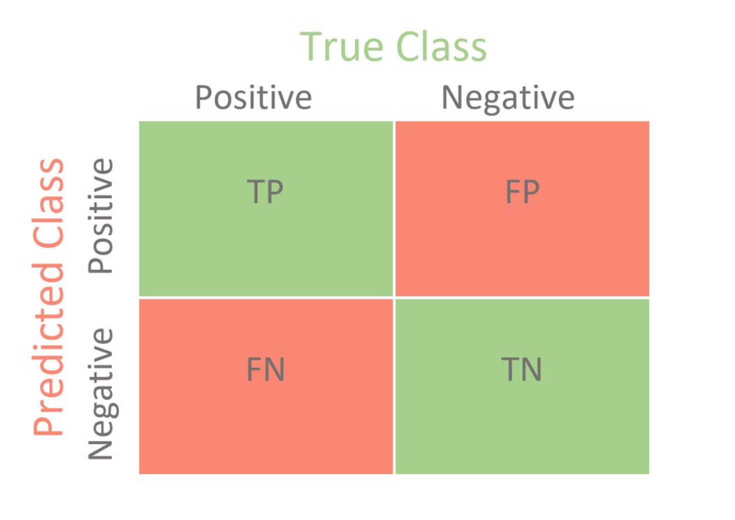
- Model Evaluation

Calculating Precision and Recall values thus minimizing the risk of False Positive (Predict Click Ads, but not click) and False Negative (Predict not click, but click ads) values. So, F1 Score is the best metrics.

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

- Yes: Customer Click on Ads (497)
- No: Customer not Click on Ads (500)

For more details, you can see all file here and code here

Data Modeling

- Modelling Result Without Transformation

Model	Accuracy	Recall	Precision	F1-Score
Log Regression	0.470000	0.000000	0.000000	0.000000
Decision Tree	0.940000	0.943396	0.943396	0.943396
Random Forest	0.956667	0.955975	0.962025	0.958991
KNN	0.630000	0.553459	0.687500	0.613240
Gradient Boost	0.966667	0.962264	0.974522	0.968354

For more details, you can see all file here and code here

Data Modeling

- Modelling Result With Transformation

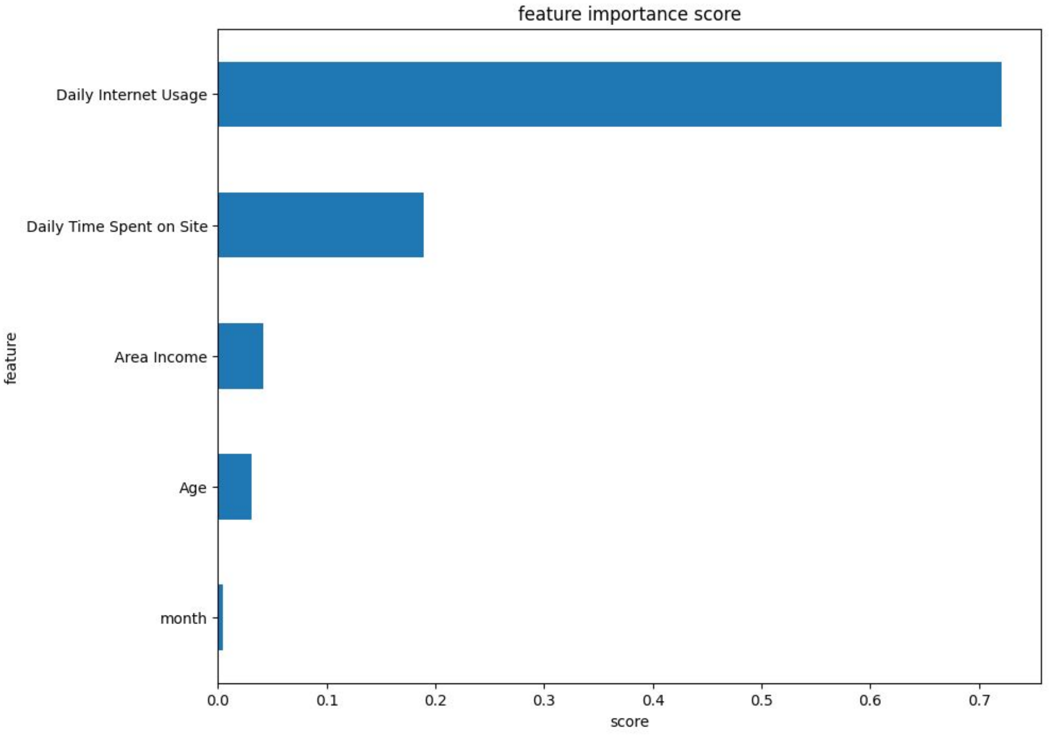
Model	Accuracy	Recall	Precision	F1-Score
Log Regression	0.960000	0.943396	0.980392	0.961538
Decision Tree	0.936667	0.943396	0.937500	0.940439
Random Forest	0.950000	0.943396	0.961538	0.952381
KNN	0.756667	0.698113	0.816176	0.752542
Gradient Boost	0.966667	0.962264	0.974522	0.968354

For more details, you can see all file here and code here

Data Modeling

- Feature Importance

Based on the feature importance scores from the machine learning models, it can be seen that the top four important features for predicting customer clicked ads are Daily Internet Usage, Daily Time Spent on Site, Area Income, and Age. Daily Internet Usage has the highest importance score across all models, followed by Daily Time Spent on Site and Area Income. Age also shows a significant importance score in predicting the outcome of the classification. This indicates that the higher the daily internet usage and daily time spent on site, the higher the likelihood of a customer clicking on an ad. Moreover, customers with higher area income and younger age are more likely to click on ads as well. Therefore, businesses can focus on these features to create a more targeted advertising strategy that will increase the likelihood of customers clicking on their ads.



For more details, you can see all file here and code here

Data Modeling

- Confusion Matrix

		Prediction	
		Click on Ads	Not Click
Actual	Click on Ads	153	6
	Not Click	4	137

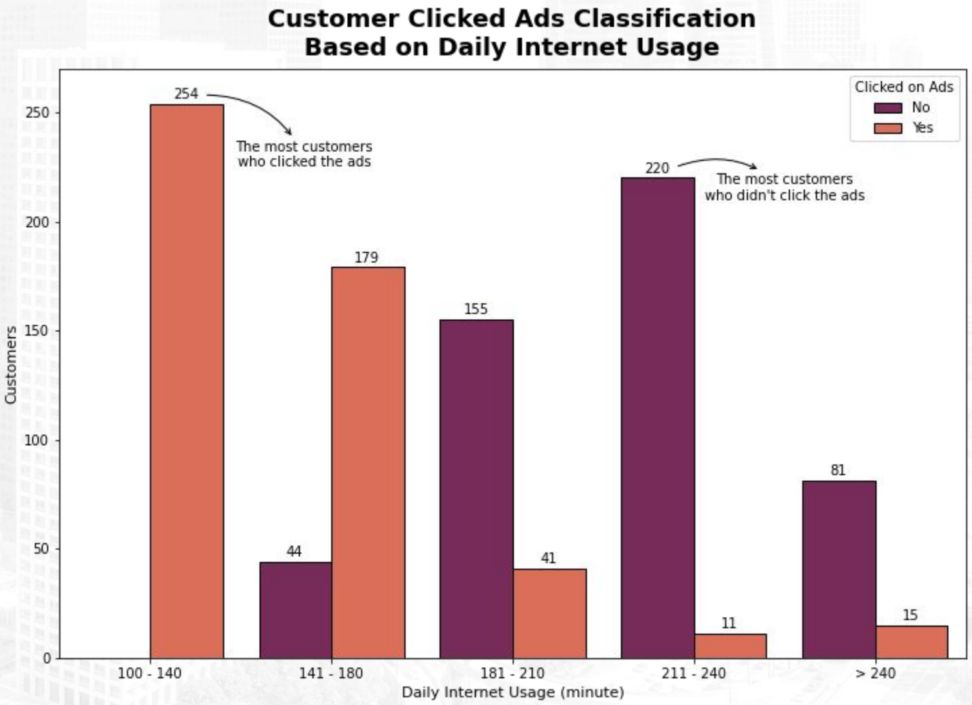
- Model Result Interpretation

From the two tables, we can see that the performance of the models with standarization and/or normalization is generally better than the models without it. The Logistic Regression model showed the biggest improvement in terms of Accuracy, Recall, Precision, and F1-Score. However, the Decision Tree model's performance decreased after normalization/standardization. The Random Forest model's performance remained almost the same. The KNN model showed a significant improvement in terms of Accuracy, Recall, Precision, and F1-Score. The Gradient Boost model's performance remained the same. Overall, it can be concluded that standarization and/or normalization improves the performance of most models and the best model based on F1 Score is Gradient Boosting. So, the model that we will use is Gradient Boosting from experiment with feature transformation.

For more details, you can see all file here and code here

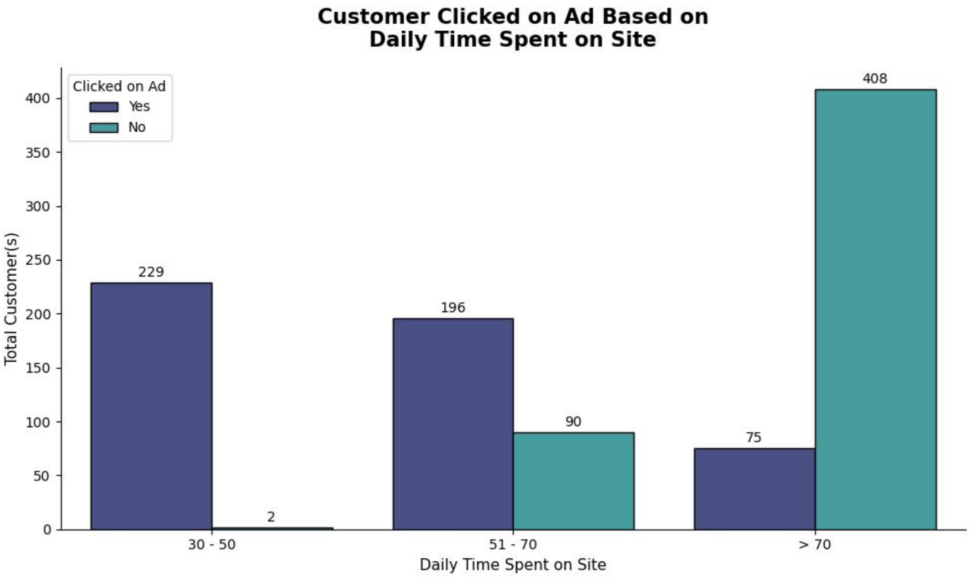
Business Recommendation & Simulation

- Target audience with a daily internet usage of 100-140 and 141-180 as they have the highest percentage of customers who clicked on the ad. This indicates that they are didn’t have time and more likely to be interested in the products or services advertised. Give them special promo or discount so they won't just click the ads but will buy our product.
- Consider improving and optimizing the ad campaign to better target customers with daily internet usage between 181-225+, as they have a high percentage of customers who did not click on the ad. This could be achieved by tailoring the ad content and targeting to better match their interests and preferences by further analysis.



For more details, you can see all file here and code here

Business Recommendation & Simulation



- Focus on targeting the audience with a daily time spent on site of 30-70, as they have the highest percentage of customers who clicked on the ad. This indicates that they are interested in the products or services advertised. Adjust the ad targeting strategies and improve the user experience as this may indicate issues with the website or advertising strategies. This could include improving the website's design and usability, providing clear and concise ad content, and ensuring that the ad placement is visible and attractive to potential customers.
- For the audience with a daily time spent on site of >70, it is important to optimize the ad campaign by tailoring the ad content and targeting to better match their interests and preferences. This could be achieved by analyzing their behavior on the website, conducting surveys or focus groups to understand their needs and preferences, and providing suitable promotional offers or products to encourage them to click and purchase. It is also important to ensure that the ad content is relevant and valuable to them, as they may be less likely to engage with ads that do not meet their needs or interests.

For more details, you can see all file here and code here

Business Recommendation & Simulation

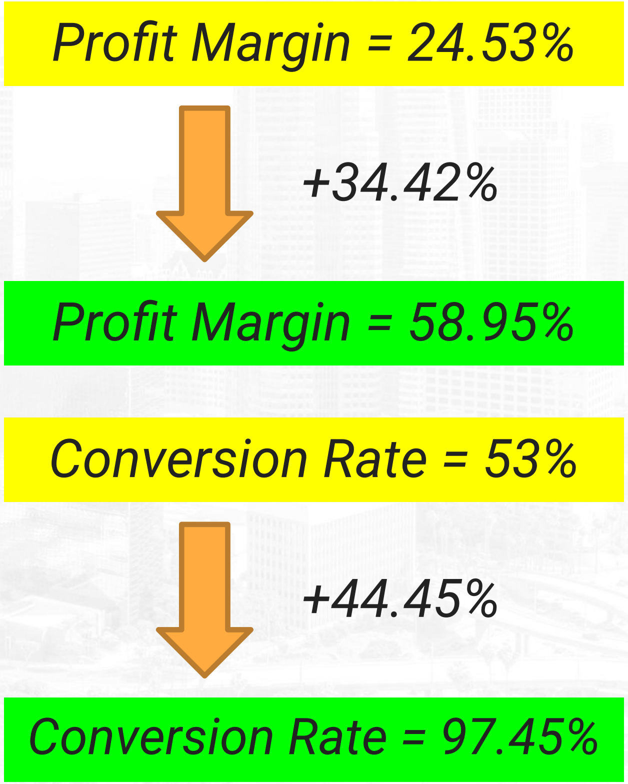
Business Simulation

- Profit Margin Without Machine Learning (300 Subject & 159 Click Ads):

Cost	300 x Rp. 2000	Rp. 600.000
Revenue	159 x Rp. 5000	Rp. 795.000
Profit	Revenue - Cost	Rp. 195.000
Profit Margin	Profit / Revenue	24.53%

- Profit Margin With Machine Learning Gradient Boosting (157 Subject & 154 Click Ads):

Cost	157 x Rp. 2000	Rp. 314.000
Revenue	154 x Rp. 5000	Rp. 765.000
Profit	Revenue - Cost	Rp. 451.000
Profit Margin	Profit / Revenue	58.95%



Profit Margin = 24.53%

+34.42%

Profit Margin = 58.95%

Conversion Rate = 53%

+44.45%

Conversion Rate = 97.45%

For more details, you can see all file here and code here

Thank You!

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