Predict Customer Clicked Ads Classification by Using Machine Learning

By: Joni Syofian

Supported by:
Rakamin Academy

About Me

Joni is a fresh graduate student from Bandung Institute of Technology. He is interested in data science, data analytics, and ocean issues. To improve his skills in the field of data, he took several courses and just completed a data science bootcamp with a good grade.

Table of Content

- Business Understanding
- Exploratory Data Analysis
- Data Preprocessing
- Modeling
- Business Recomendation and Business Simulation

BUSINESS UNDERSTANDING

Background

A company in Indonesia wants to know the effectiveness of an advertisement that they display. It is important for the company to be able to find out how much the advertisement has been marketed so that it can attract customers to see the advertisement.By processing historical advertisement data and finding insights and patterns that occur, it can help companies determine marketing targets. The focus of this case is to create a machine learning classification model that functions to determine the right target customers.

Problem

The business team wants to **optimize their advertising** methods on digital platforms in order to get potential users to click on a product so that **the costs to be incurred are not too large**.

Goal

Make target marketing effective by using machine learning so that it can increase the click-through rate (CTR) and reduce costs incurred.

Objective

Building supervised machine learning to classify the right customers for marketing targets .

Business Metrics

- Click Through Rate (CTR)
- Total Cost

Data Overview

Data Overview

```
<class 'pandas.core.frame.DataFrame'>
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)
memory usage: 86.1+ KB
```

Description

This is historical advertisement data related to customers.

Rows and columns

It consists of 1000 rows and 11 columns.

Missing Values

There are missing values in the columns `Daily Time Spent on Site`, `Area Income`, `Daily Internet Usage`, and "Male`.

Duplicates Data

There are no duplicate data.

Exploratory Data Analysis

Descriptive Statistics

Numerical Data

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

- There are different mean and median values for each numerical column, which indicate that the data distribution is not normal, but this will be confirmed again by the distribution plot and also the skewness values.
- There is a significant difference between the minimum value in the lower quartile (25%) and the maximum value in the upper quartile (75%), which indicates that there are outliers in the data, but this will be confirmed later using box plots.
- A large enough standard deviation value indicates that the data is spread out (far from the average).
- The average customer spends around 64 minutes on the platform or website.
- The average customer age is 36 years old.
- The average income area is IDR 384,864,700
- The average customer spends 179 minutes per day on the internet.

Descriptive Statistics

Categorical Data

	Male	Clicked on Ad	city	province	category
count	997	1000	1000	1000	1000
unique	2	2	30	16	10
top	Perempuan	No	Surabaya	Daerah Khusus Ibukota Jakarta	Otomotif
freq	518	500	64	253	112

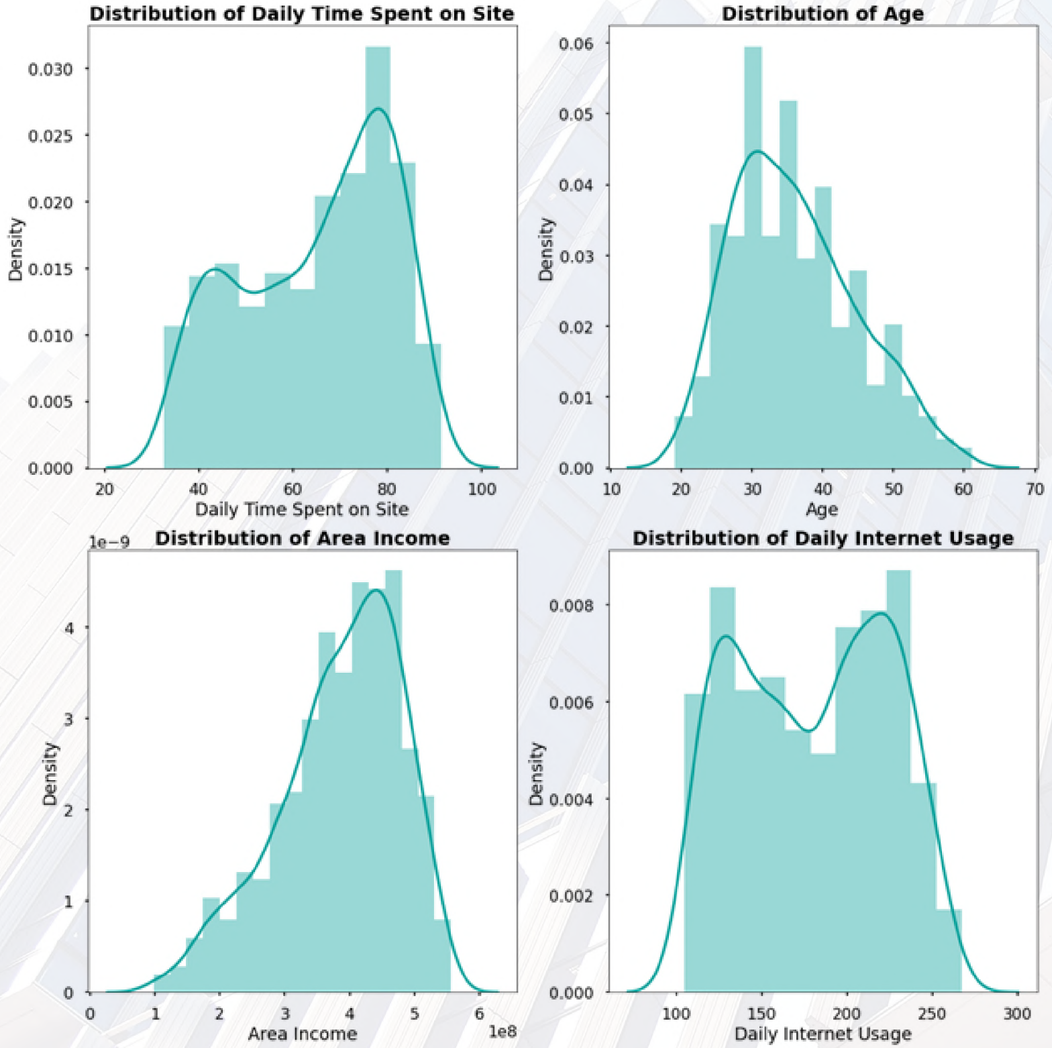
- In the male column, the most valuable word is "Perempuan", so most of the customers are women. The naming of this column is a bit confusing, so we will change the column name later.
- Clicked on Ad both yes and no values have the same value.
- The most populous cities are Surabaya and Bandung.
- The province with the most customers is the province of the Special Capital Region of Jakarta.
- Based on "category," the most valuable is "Otomotif".

Datetime Data

	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

- There is a difference between the "count" and "unique" values, which indicate that there are customers who click on ads simultaneously.
- On 2016-05-26 15:40:00, two different customers clicked on ads at the same time.
- The advertisement was clicked for the first time on 2016-01-01 02:52:00.
- The advertisement was most recently clicked on 2016-07-24 00:22:00.so that from the data, it was found that the advertisement lasted for 7 months, namely from January 2016 to July 2016.

Univariate Analysis



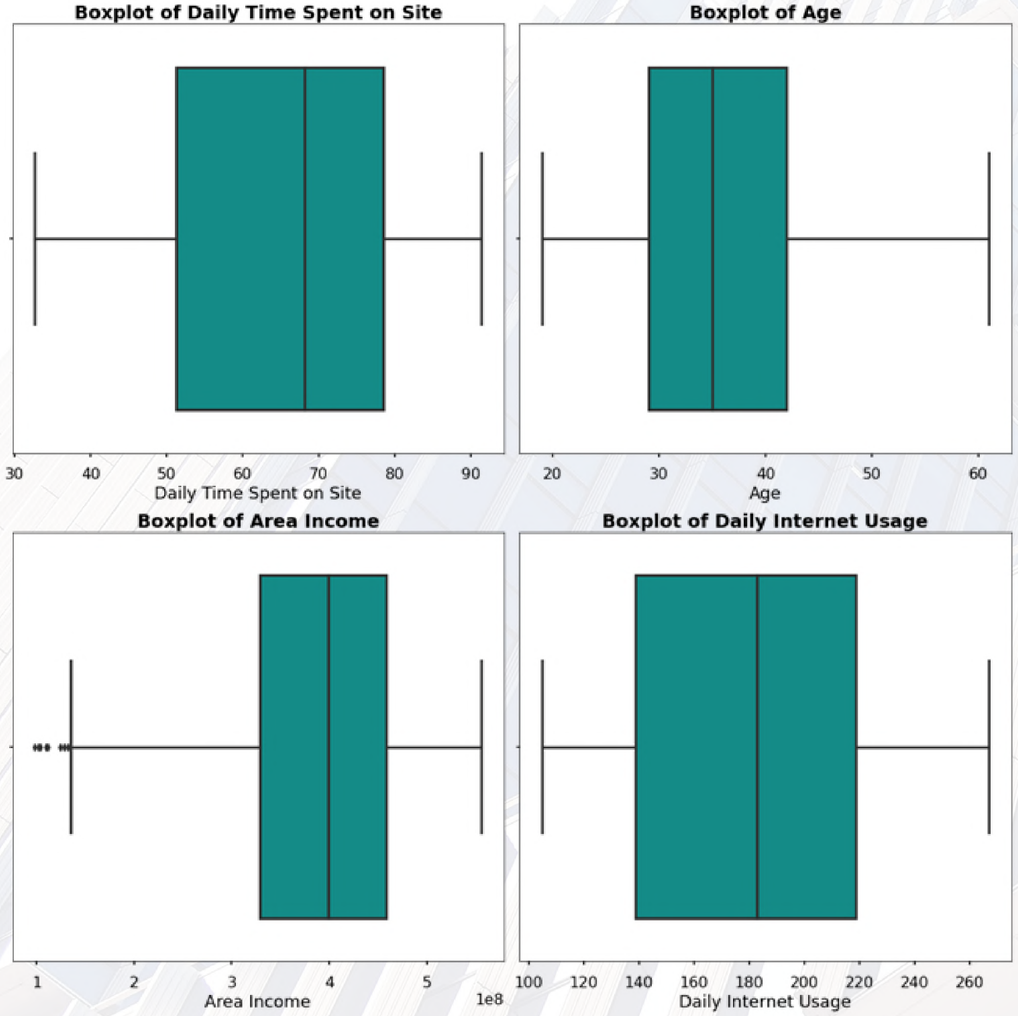
Numerical Data

- Customers spend the most time on the site, ranging from 75 - 85 minutes.
- The majority of customers are between the ages of 30 and 35.
- Based on income area, the highest area of income is in the range of IDR 400,000,000 - IDR 500,000,000.
- Customers use the internet the most during the day between 220-240 minutes.

```
Skewness Daily Time Spent on Site : -0.36975576201210597
Skewness Age : 0.4791416884125751
Skewness Area Income : -0.6443017726963709
Skewness Daily Internet Usage : -0.03139523418033974
```

From the distribution plot and skewness values, it can be seen that the numerical data has a distribution close to the normal distribution, with 3 out of 4 numeric variables having negative skewness values ('Daily Time Spent', 'Area Income', 'Daily Internet Usage'); the only positive value is 'age'.

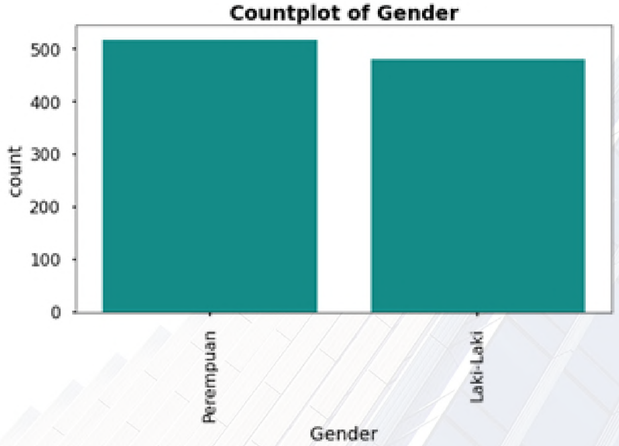
Univariate Analysis

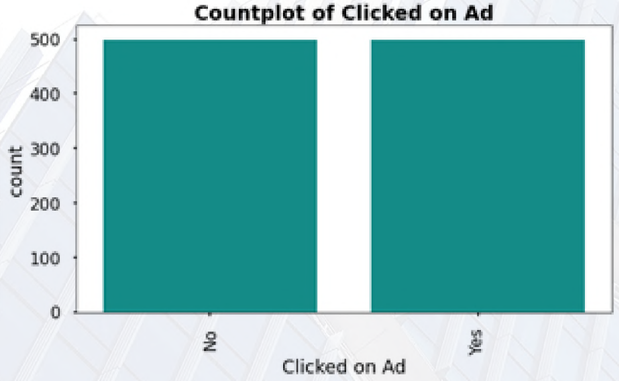


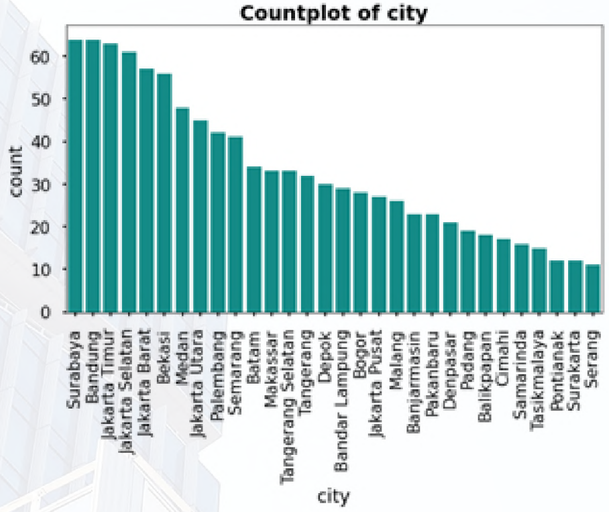
Numerical Data

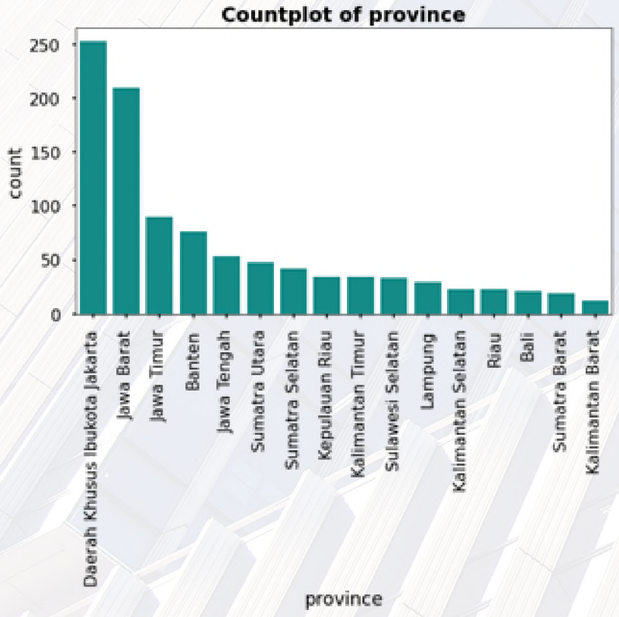
From the boxplot, it is clear that only the `Area Income` column has outliers. Later, this will be done to handle outliers.

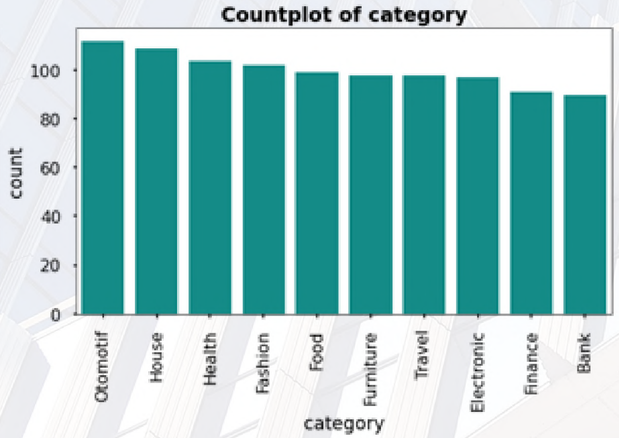
Univariate Analysis







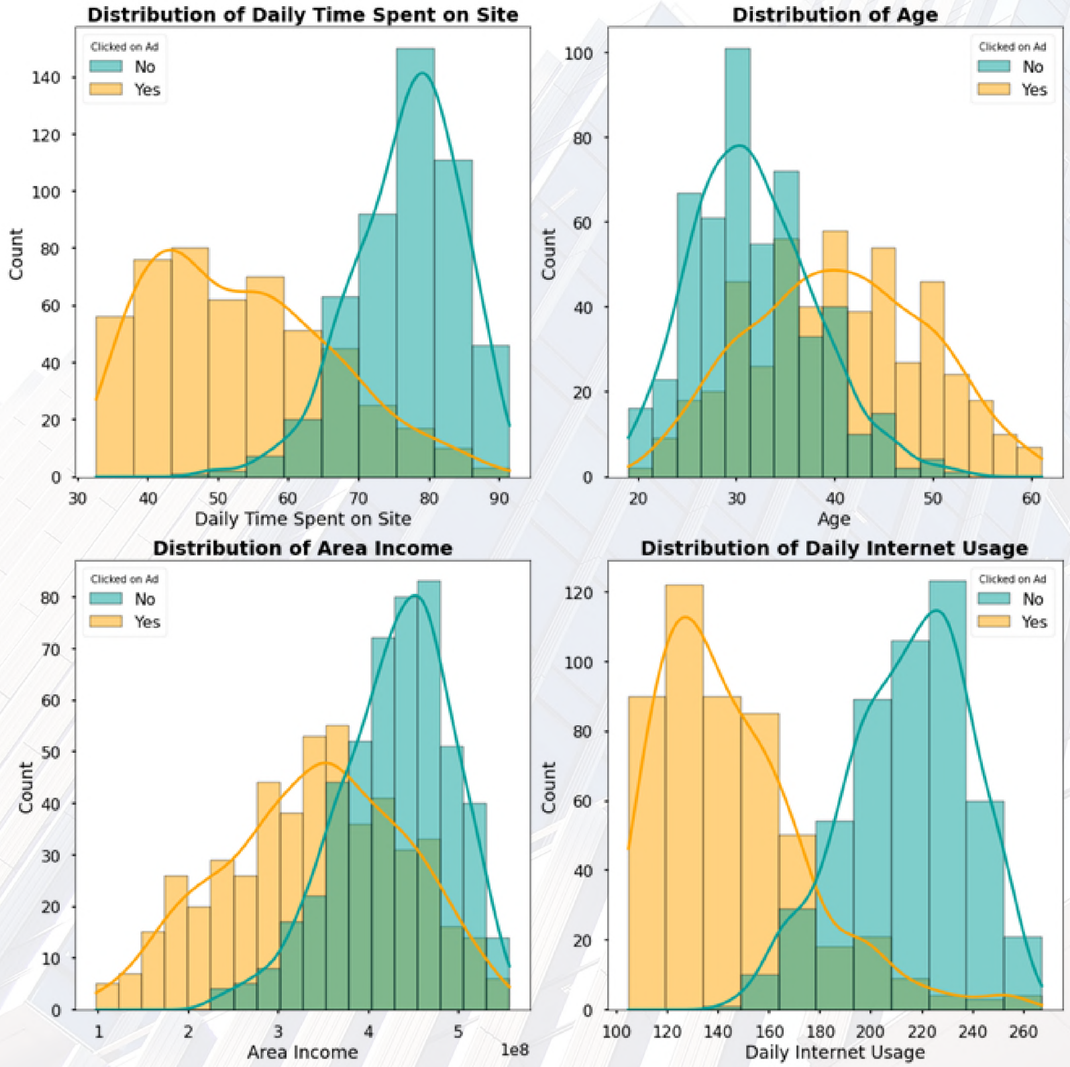




Categorical Data

- Women are the most frequent customers, although there is no significant difference.
- There is no difference in the number of customers who click on ads and those who do not click on ads.
- Bandung and Surabaya are the cities with the most customers.
- The province with the greatest number of customers is DKI Jakarata.
- "Otomotif" is the largest "category", but there is no discernible difference between it and the other "categories".

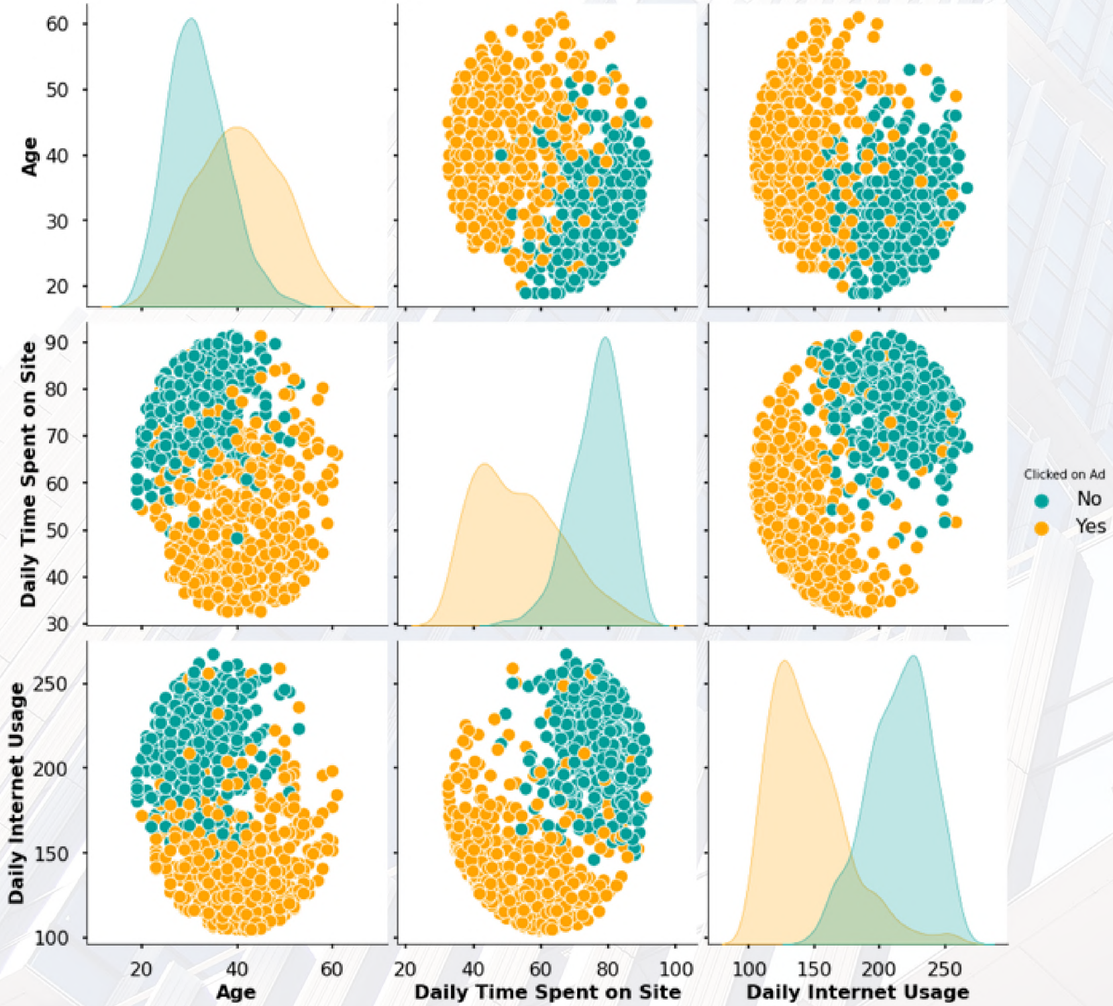
Bivariate Analysis



Numerical Data

- Customers who click on ads spend an average of 40 to 50 minutes per day on the site. while those who did not click on ads were in the range of 75 to 85 minutes.
- The age distribution of customers who click on ads resembles a normal distribution, with the majority of customers being between the ages of 35 and 45. Whereas those who did not click on ads had a skewed distribution, with the most customers being in the age range of 28–32 years.
- The income area for customers who click on ads has a distribution resembling a normal distribution, with the majority of customers having an income area of IDR 300 million to IDR 400 million. Whereas those who do not click on ads have a skewed distribution, with the most customers having an income of around IDR 430 million - IDR 480 million.
- Daily internet usage for customers who click on ads tends to be lower, with most customers having daily internet usage of 120 -140 minutes. Meanwhile, customers who do not click on ads tend to have large daily usage, with the most customers in the range of 220 to 240 minutes.

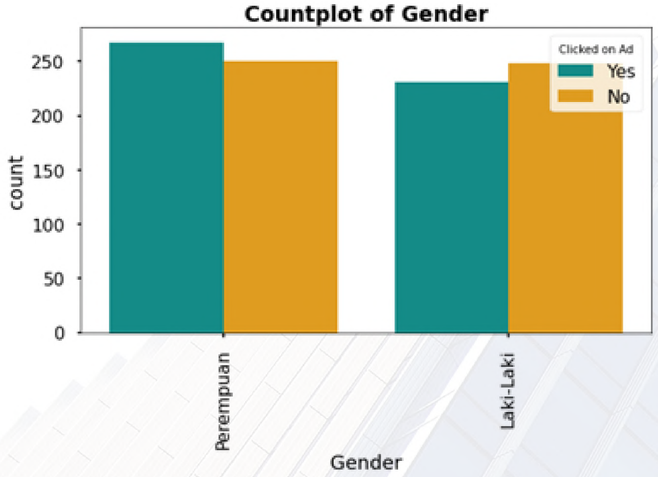
Bivariate Analysis

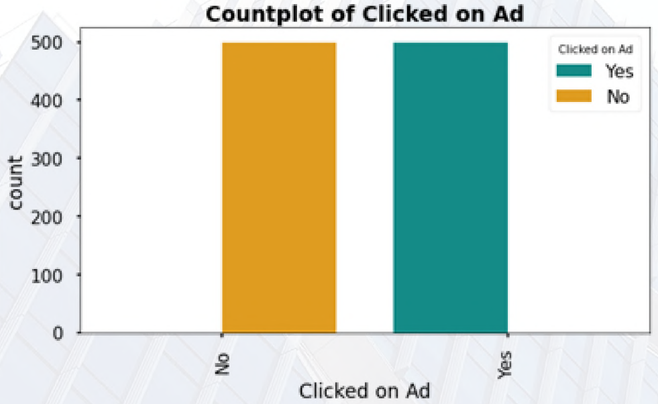


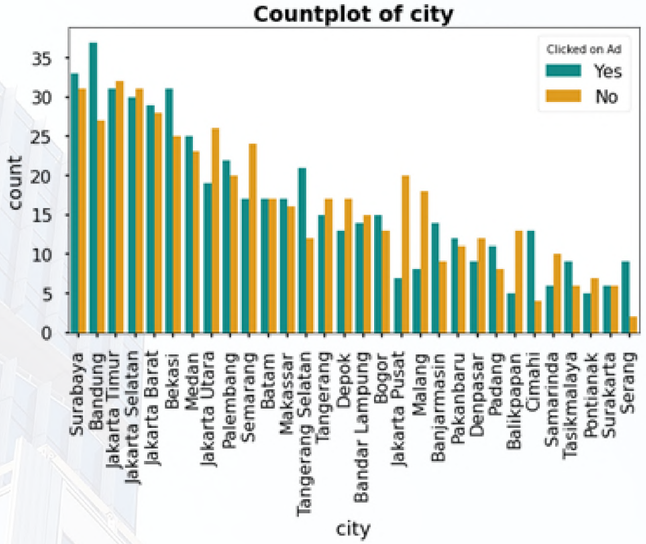
Numerical Data

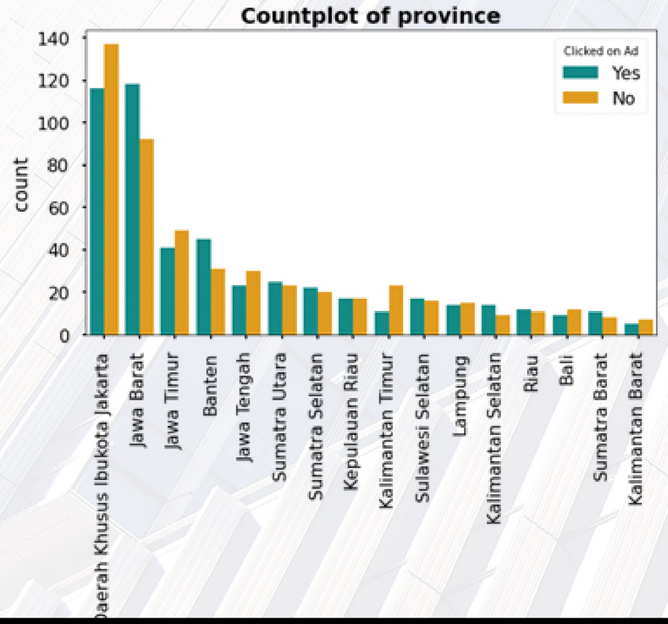
- The greater the customer's age, the lower the daily internet usage value as well as the daily time spent on site (inversely proportional).
- It was also found that the tendency of customers to click on advertisements consisted of people of various ages who had daily internet usage and also spent daily time on sites that were small.
- For daily time spent on site and daily internal usage, there is a positive correlation where the greater the daily time spent on site, the greater the daily internet usage.

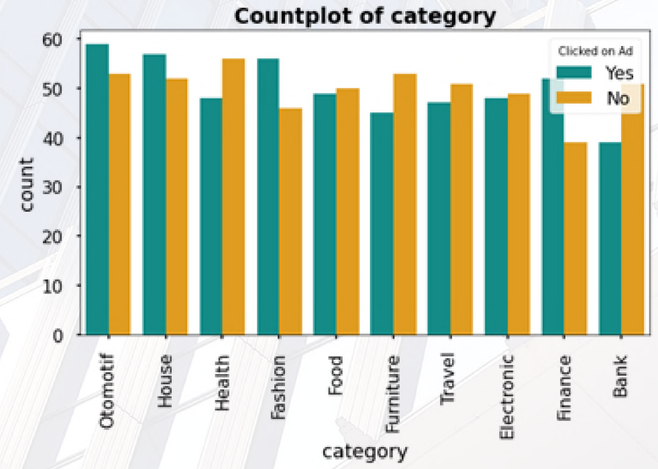
Bivariate Analysis







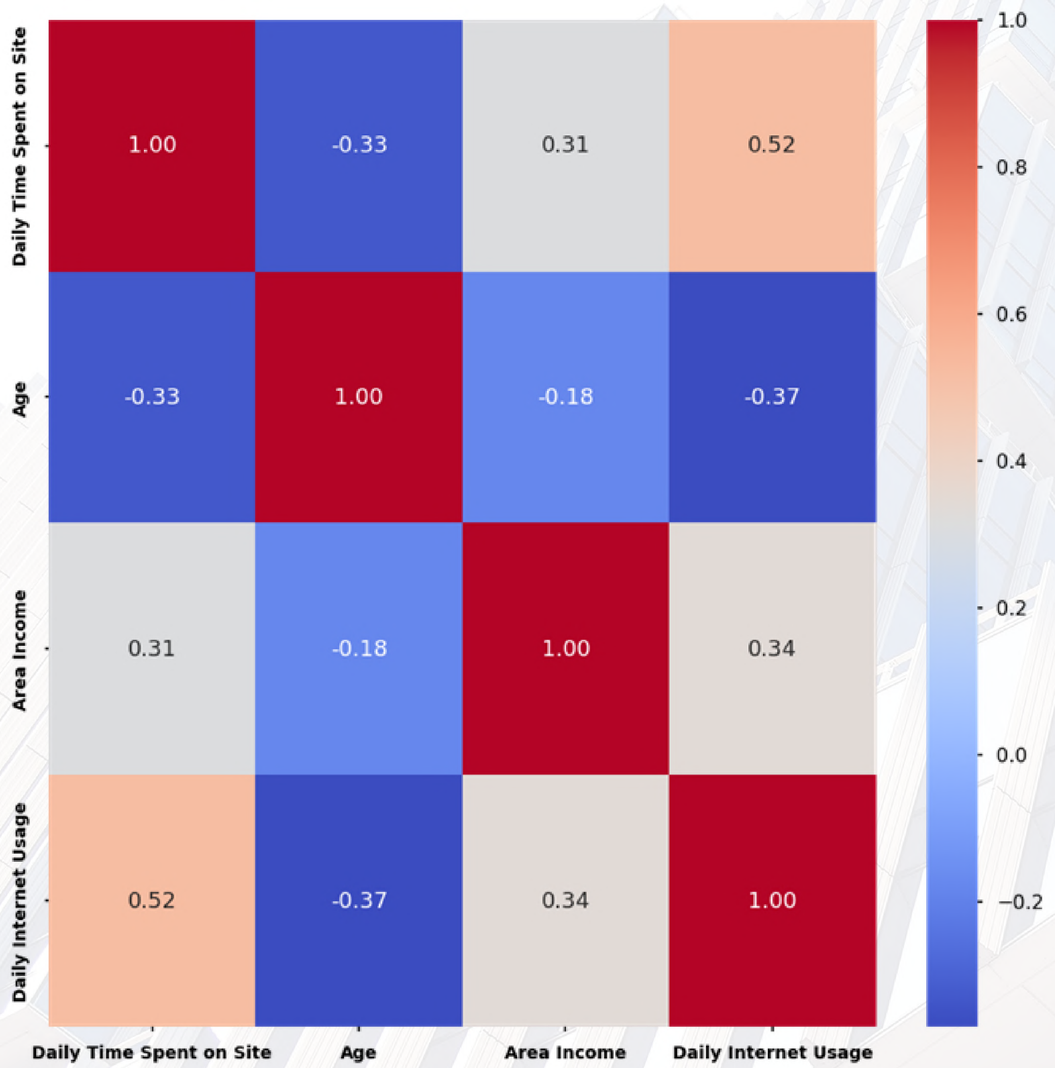




Categorical Data

- More women click on advertisements than men. On the other hand, more men do not click on ads than do click on ads. But overall, there is no significant difference.
- For the other columns, there are no values that are too dominant, and no conclusions can be drawn.

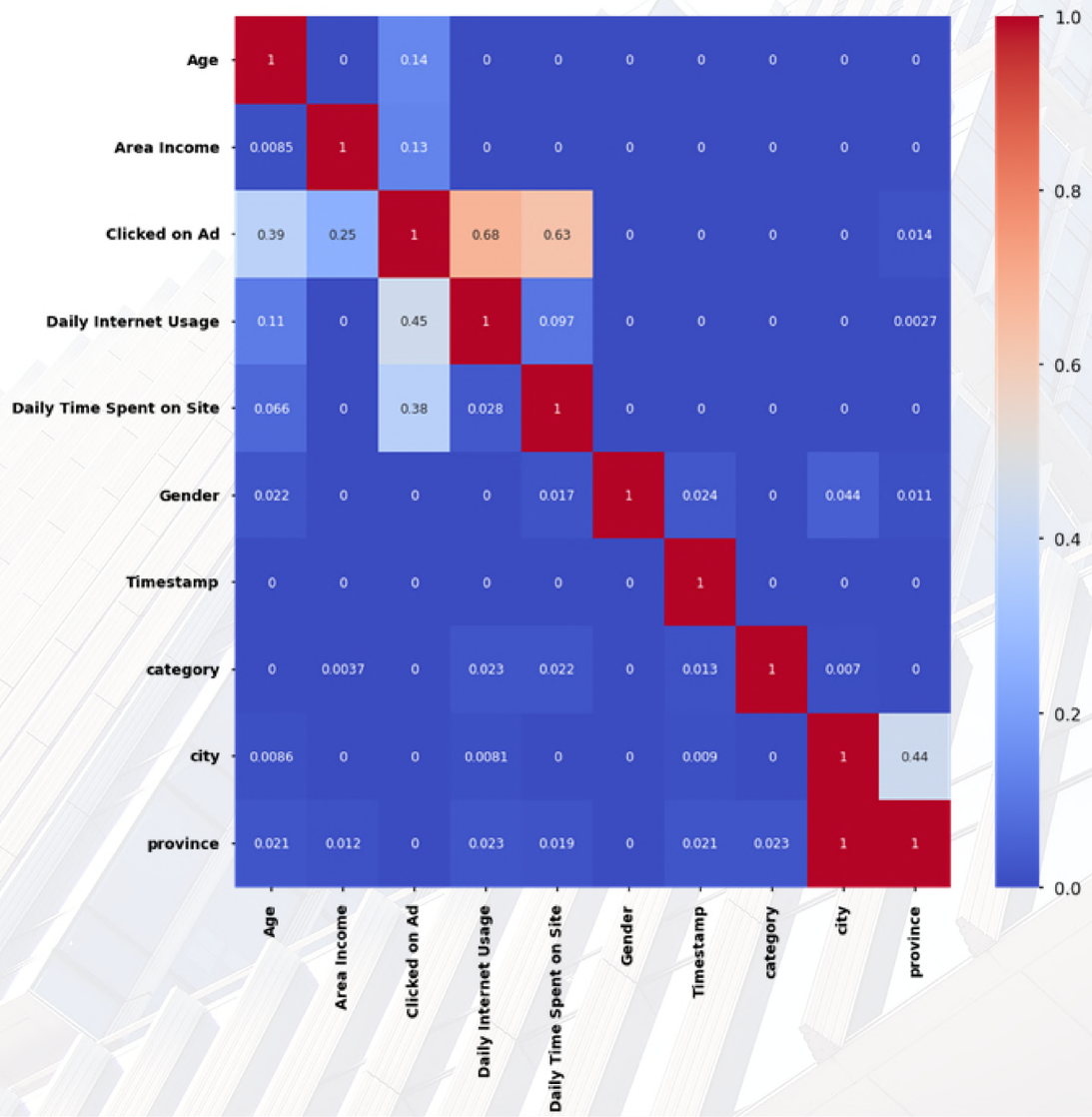
Multivariate Analysis



Correlation (Pearson)

From the heatmap correlation, it is known that the feature that has the highest correlation is between the daily time spent on site and daily internet usage, which is equal to 0.52. This number is not too large, so it does not indicate that there are redundant features, and all features can be used for machine learning. In addition, negative correlation values were also found, namely for age and daily internet usage, age with income area, and age with daily time spent on site. However, the correlation value is not large and does not exceed 0.7.

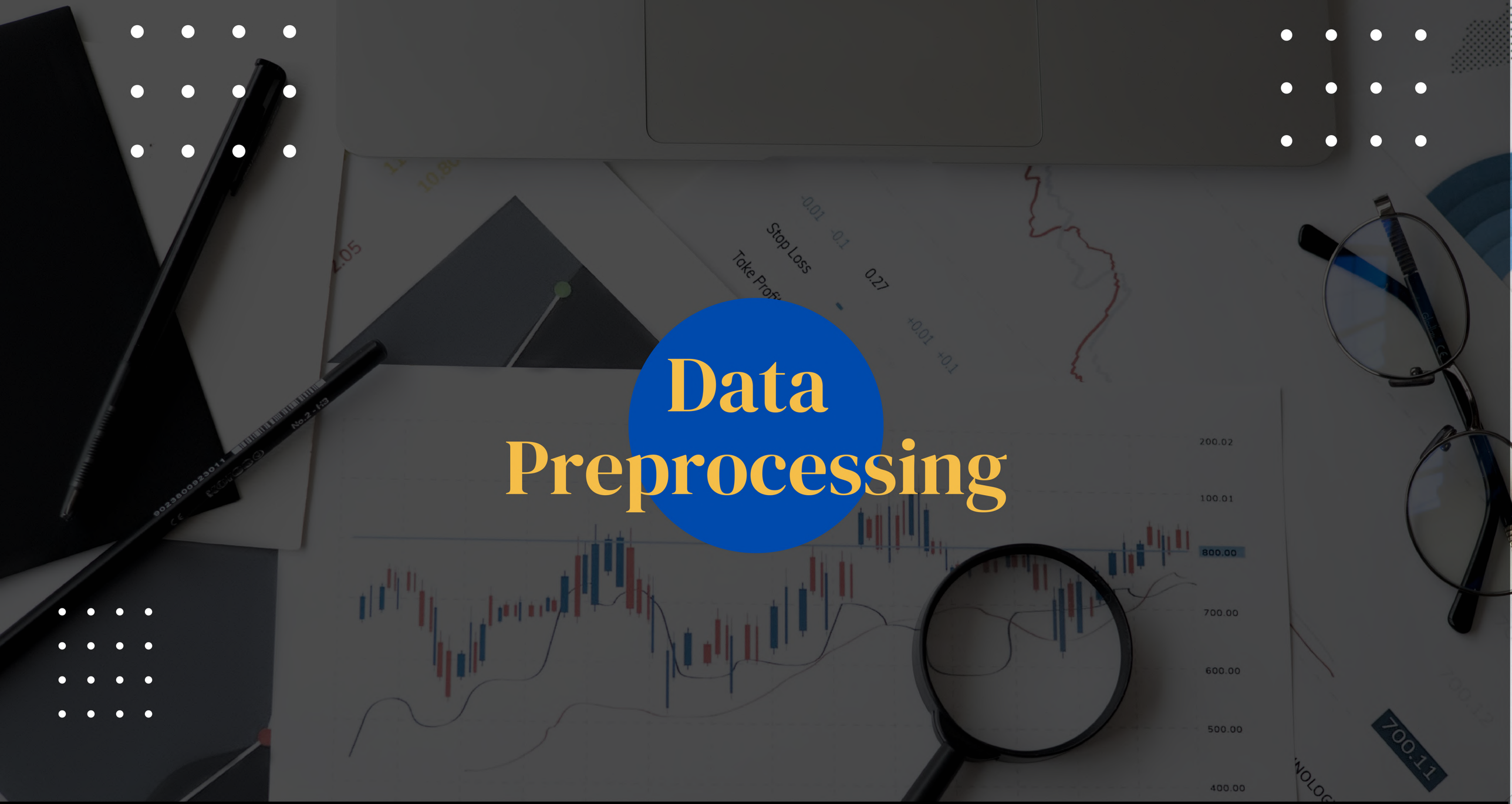
Multivariate Analysis



PPS (Predictive Power Score)

The heatmap correlation (Pearson correlation) only shows the correlation of numerical features. Meanwhile, the correlation value of the category features is not known. For that, we can use PPS (Predictive Power Score), which can see the relationship of all features with the target ("Clicked on Ad").

PPS (Predictive Power Score) found that the features `Daily Internet Usage`, `Daily Time Spent on Site`, `Age`, and `Area Income` have a high score compared to other features.



Data Preprocessing

Data Preprocessing

Define Target

The target feature in this machine learning is `Clicked on Ad` which indicates whether the customer clicked on the ad or not. Since the values for this feature are 'Yes' and 'No' we will replace them with values 1 and 0 with the name `Target`.

Handle Missing Value

For missing value handles in the four columns, it will be carried out based on the value of skewness and the highest value:

- The `Daily Time Spent on Site` column will be filled with average values due to skew symmetry.
- The `Area Income` column will be filled with the median due to skew.
- The `Daily Internet Usage` column will be filled with an average due to the symmetry skew.
- The `Gender` field will be filled with the mode.

Feature Extraction

At this stage, we created a new column based on the `Timestamp` column. The columns to create are the `month`, `days`, and `Weekdays` columns, where 0 denotes Monday and 6 denotes a week. Furthermore, it will be grouped into weekdays and weekends (Saturday and Sunday).

Handling Outliers

As it is known that in the column 'Area Income' there are outliers, we will handle this by removing the outliers. We will be used the IQR methods. Because the data is cleaner than outliers and not much data is removed (only 0.3% removed).

Feature Selection

At this stage, columns are dropped that are not related to the model to be built.

Data Preprocessing

Feature Scaling and Transformation

- **Feature Encoding**

  The feature encoding technique used is one hot encoding because all categorical variables are not ordinal or multilevel variables.

- **Feature Scaling**

  1. Numerical Data

     Feature scaling is performed on the `Daily Time Spent on Site`, `Age`, and `Daily Internet Usage` features using standardization.

  2. Cyclic Data

     performed on `Month` and `Days` data by changing using the Cos and Sin functions.

Split Feature dan Target

- Target : Column `Target`
- Features : 22 other columns.

#	Column	Non-Null Count	Dtype
0	Daily Time Spent on Site	991 non-null	float64
1	Age	991 non-null	float64
2	Daily Internet Usage	991 non-null	float64
3	Area Income	991 non-null	float64
4	Target	991 non-null	float64
5	Gender_Laki-Laki	991 non-null	float64
6	Gender_Perempuan	991 non-null	float64
7	category_Bank	991 non-null	float64
8	category_Electronic	991 non-null	float64
9	category_Fashion	991 non-null	float64
10	category_Finance	991 non-null	float64
11	category_Food	991 non-null	float64
12	category_Furniture	991 non-null	float64
13	category_Health	991 non-null	float64
14	category_House	991 non-null	float64
15	category_Otomotif	991 non-null	float64
16	category_Travel	991 non-null	float64
17	Types of days_Weekday	991 non-null	float64
18	Types of days_Weekend	991 non-null	float64
19	Sin_Month	991 non-null	float64
20	Cos_Month	991 non-null	float64
21	Sin_Days	991 non-null	float64
22	Cos_Days	991 non-null	float64

dtypes: float64(23)

Columns used to build machine learning


Stop Loss
0.27
200.02
100.01
800.00
700.00
600.00
500.00
400.00
700.11

Modeling

Modeling

Before doing modeling, a split-tran test (70:30) data set is first performed. Then two schemes are carried out, namely modeling without feature scaling and modeling with feature scaling.

Features used in models without feature scaling

#	Column	Non-Null Count	Dtype
0	Daily Time Spent on Site	693 non-null	float64
1	Age	693 non-null	int64
2	Area Income	693 non-null	float64
3	Daily Internet Usage	693 non-null	float64
4	Month	693 non-null	int64
5	Days	693 non-null	int64
6	Gender_Laki-Laki	693 non-null	uint8
7	Gender_Perempuan	693 non-null	uint8
8	category_Bank	693 non-null	uint8
9	category_Electronic	693 non-null	uint8
10	category_Fashion	693 non-null	uint8
11	category_Finance	693 non-null	uint8
12	category_Food	693 non-null	uint8
13	category_Furniture	693 non-null	uint8
14	category_Health	693 non-null	uint8
15	category_House	693 non-null	uint8
16	category_Otomotif	693 non-null	uint8
17	category_Travel	693 non-null	uint8
18	Types of days_Weekday	693 non-null	uint8
19	Types of days_Weekend	693 non-null	uint8

dtypes: float64(3), int64(3), uint8(14)

Features used in models with feature scaling

#	Column	Non-Null Count	Dtype
0	Gender_Laki-Laki	693 non-null	uint8
1	Gender_Perempuan	693 non-null	uint8
2	category_Bank	693 non-null	uint8
3	category_Electronic	693 non-null	uint8
4	category_Fashion	693 non-null	uint8
5	category_Finance	693 non-null	uint8
6	category_Food	693 non-null	uint8
7	category_Furniture	693 non-null	uint8
8	category_Health	693 non-null	uint8
9	category_House	693 non-null	uint8
10	category_Otomotif	693 non-null	uint8
11	category_Travel	693 non-null	uint8
12	Types of days_Weekday	693 non-null	uint8
13	Types of days_Weekend	693 non-null	uint8
14	norm_Daily Time Spent on Site	693 non-null	float64
15	norm_Age	693 non-null	float64
16	norm_Daily Internet Usage	693 non-null	float64
17	norm_Area Income	693 non-null	float64
18	Sin_Month	693 non-null	float64
19	Cos_Month	693 non-null	float64
20	Sin_Days	693 non-null	float64
21	Cos_Days	693 non-null	float64

dtypes: float64(8), uint8(14)

Modeling

Features used in models without feature scaling (Model Evaluation)

	MLA used	Train Accuracy (%)	Test Accuracy (%)	Precission (%)	Recall (%)	F1-Score (%)	roc-auc (test prob)	roc-auc (train prob)	roc-auc (crossval train)	roc-auc (crossval test)	Time used
0	LogisticRegression	49.9	51.7	0.0	0.0	0.0	0.713	0.787	0.765	0.765	0.336980
1	KNeighborsClassifier	76.5	65.8	68.1	54.9	60.8	0.681	0.856	0.848	0.702	0.687961
2	DecisionTreeClassifier	100.0	94.0	93.1	93.8	93.4	0.946	1.000	1.000	0.931	0.157992
3	RandomForestClassifier	100.0	97.3	98.6	95.8	97.2	0.991	1.000	1.000	0.991	2.541858
4	GradientBoostingClassifier	99.9	97.3	97.2	97.2	97.2	0.990	1.000	1.000	0.987	2.058881
5	XGBClassifier	100.0	96.3	97.2	95.1	96.1	0.991	1.000	1.000	0.989	1.189932

it is found that **Random Forest** is an algorithm that has better performance than other algorithms. However, in terms of **computational time**, **random forest** is the most **time-consuming** algorithm. While the algorithm that requires **the least time is the Decision Tree algorithm.**

Features used in models with feature scaling (Model Evaluation)

	MLA used	Train Accuracy (%)	Test Accuracy (%)	Precission (%)	Recall (%)	F1-Score (%)	roc-auc (test prob)	roc-auc (train prob)	roc-auc (crossval train)	roc-auc (crossval test)	Time used
0	LogisticRegression	96.0	98.7	100.0	97.2	98.6	0.993	0.992	0.765	0.765	0.417974
1	KNeighborsClassifier	86.3	73.8	73.2	72.2	72.7	0.790	0.927	0.848	0.702	0.515970
2	DecisionTreeClassifier	100.0	92.6	94.2	91.0	92.6	0.933	1.000	1.000	0.932	0.158023
3	RandomForestClassifier	100.0	98.0	98.6	97.2	97.9	0.992	1.000	1.000	0.990	2.523826
4	GradientBoostingClassifier	99.9	97.3	97.2	97.2	97.2	0.992	1.000	1.000	0.987	2.089882
5	XGBClassifier	100.0	96.6	97.2	95.8	96.5	0.990	1.000	1.000	0.989	1.138934

it is found that the **model performance of various algorithms had increased. which has experienced a drastic increase is the logistic regression algorithm**, which has a precision of 100%, a recall of 97.2%, and an F1-Score of 98.6%, as well as a model that has a good evaluation matrix value and a short computation time. **However, the cross validation value of the ROC-AUC model is the lowest compared to the other models.**

The **model's goal** is to **predict the maximum number of potential customers who will click on an ad.** Therefore we must **minimize False Positives** where customers who do not click on ads are predicted to click on ads wrongly. This will lead to retargeting the wrong market and then cause potential losses because we have spent marketing costs on the wrong target.

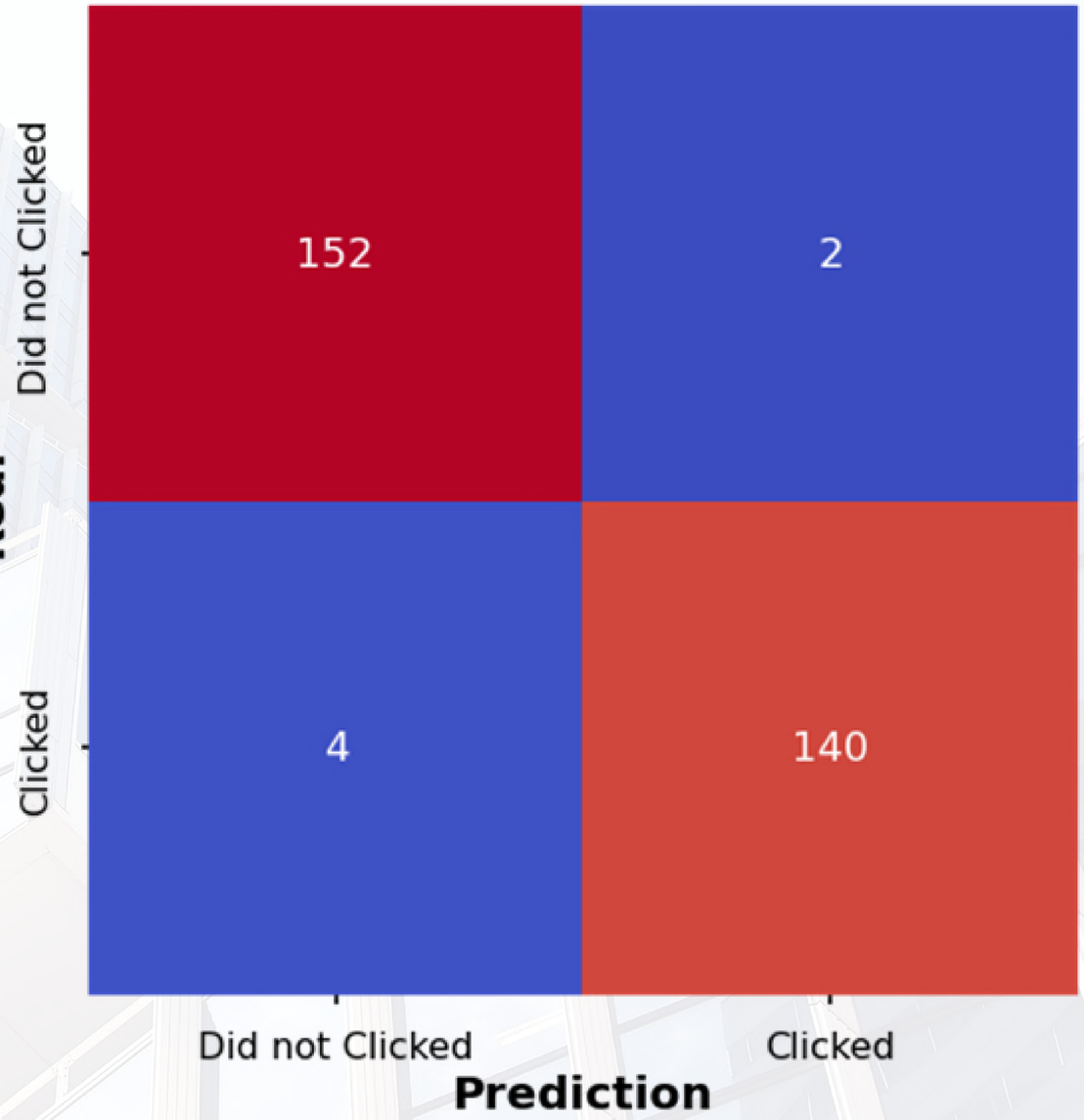
Therefore, **we have to optimize the precision score** while still considering other metrics to be at their maximum.

Finally, the **model chosen** is the one with the **Random Forest algorithm** because **it has a greater precession and accuracy value than the other models.** However, we can also use a model with the gradient boosting algorithm, which is not much different from the results of the Random Forest algorithm model.

Confusion Matrix

Confusion matrix of the random forest algorithm model with a modeling scheme with feature scaling

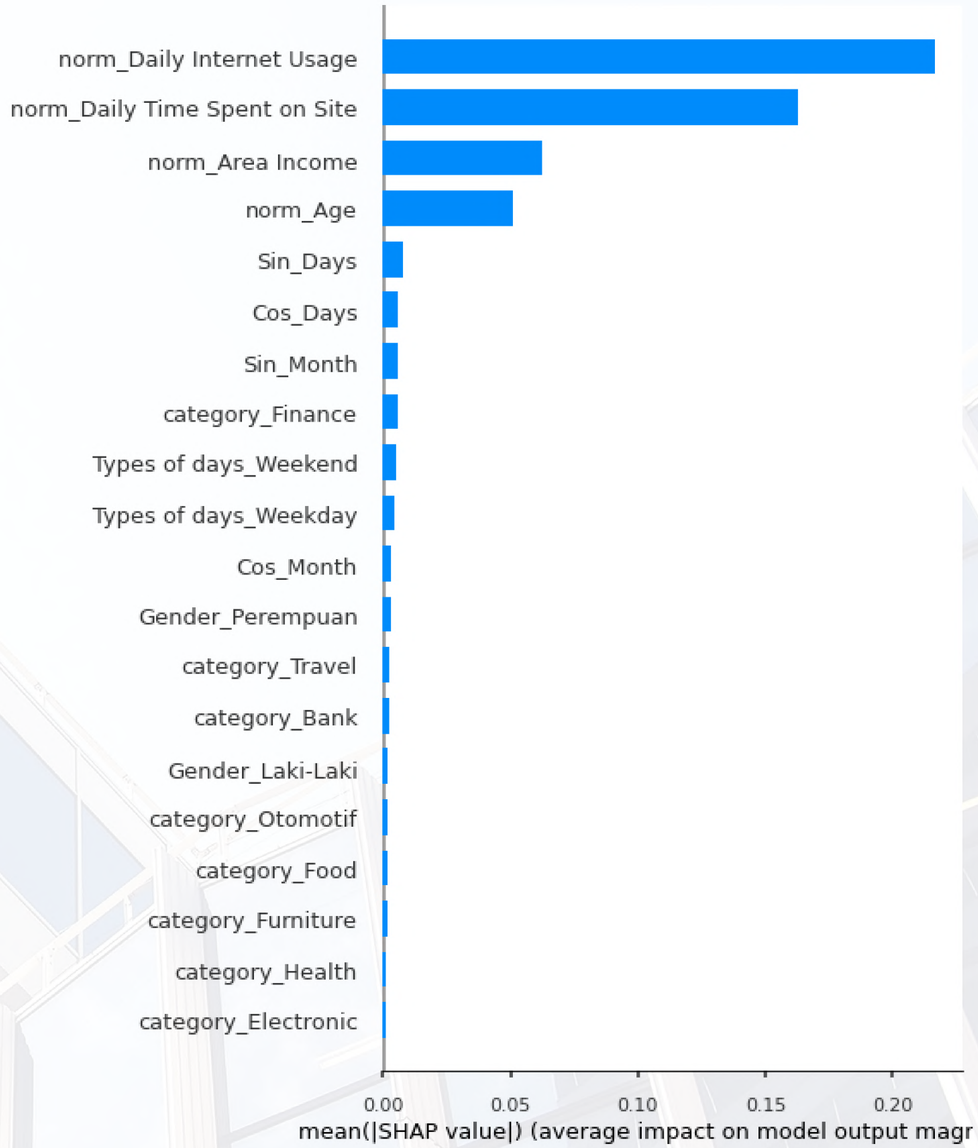
From the confusion matrix, it can be seen that the model with the random forest algorithm has good performance, where the predicted number of clicks but no clicks (false positives) is of little value.



Feature Importance

Feature Importance of Model with Random Forest Algorithm (with feature scaling)

From the model that has been built, the features that have a major influence on the model are obtained. The are 4 important features are `Daily Internet Usage`, `Daily Spent on Site`, `Area Income` and `Age` (picture below). These four features greatly affect the model that the customer will click on the ad or not. This important feature will be used as a benchmark for business recommendations.



Business Recommendation & Business Simulation

Business Recommendation

Based on EDA and feature importance, business recommendations that can be submitted are as follows:

1. **Customers who click on ads** are customers who **spend time on the internet and websites in a short time.** Therefore, the thing that can be done is to **optimize advertisements with more attractive packaging** so that when a customer enters the website, they immediately click on the advertisement. On the other hand, adding advertisements is deemed unnecessary because it can increase costs and customers who click on advertisements do not spend much time on the internet.

2. **Based on "Area Income"**, customers are **normally distributed**, which means that customers who click on ads are those **with small to the largest income areas**. The action that can be taken is to **make sure that the advertisements given are advertisements that can cover all income groups** so that both customers with low and high incomes are interested in clicking on the advertisements given.

3. **Based on "Age"**, customers are **normally distributed**, which means that customers who click on ads range in age from **young to old. The action** that can be taken is **to ensure that the advertisements given are those that can cover all age groups** so that both young and old customers are interested in clicking on the advertisements given.

Business Simulation

The data used in this business simulation is the data from the split_train_test (data test). The details are as follows:

```
Target
0          154
1          144
dtype: int64
```

0 : Who did not click on the ad
1 : Who clicked on the ad

It can be seen that the number of customers is 298 customers. There are two schemes in this business simulation, namely without machine learning and with machine learning. With the assumptions used are as follows:

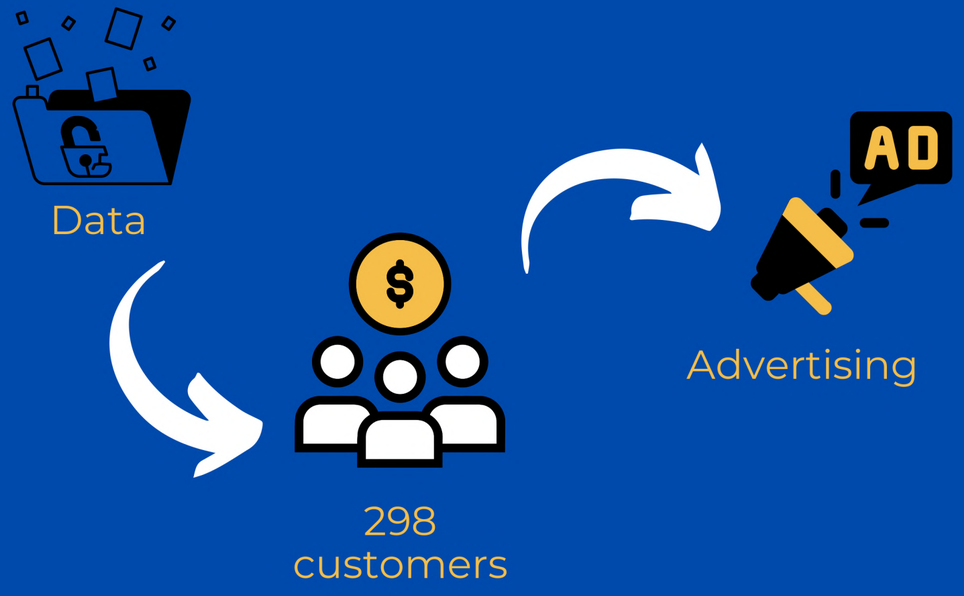
- Advertising costs per customer = Rp. 1.000
- Profit earned when a customer clicks on an ad = Rp. 5.000

Formulas used	
Click Through Rate (CTR)	= (Number of customers who clicked : total number of customers) *100%
Total Cost	= Advertising costs per customer * Number of customers given the advertisement
Revenue	= Profit earned when a customer clicks on an ad * Number of customers who clicked
Profit	= Revenue – Total Cost

Business Simulation

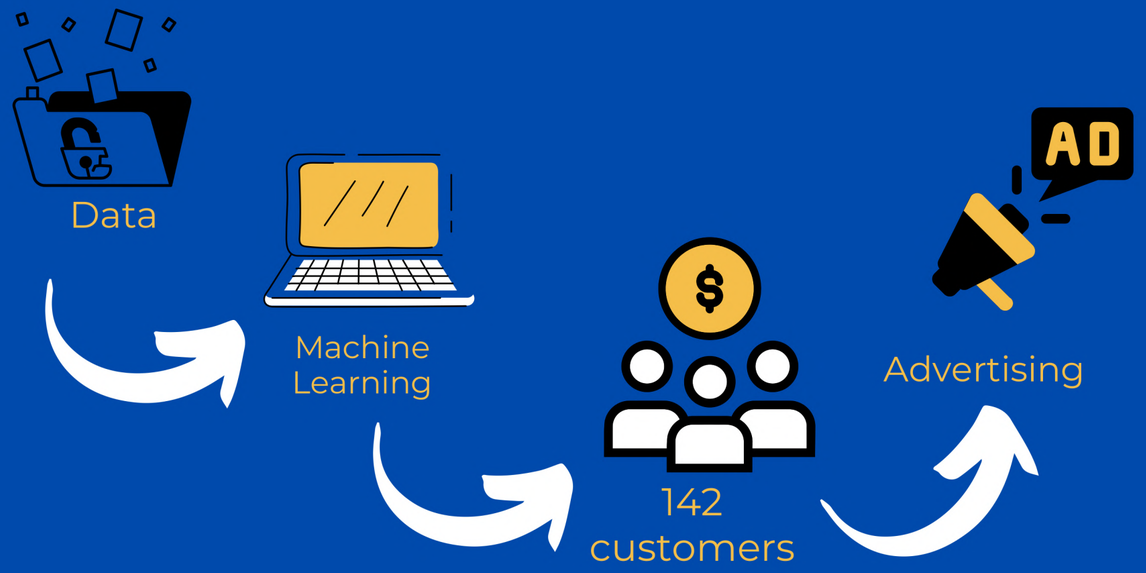
The following is an overview of the two business simulation schemes.

Without Modeling



In this scheme, advertising is applied to all 298 customers

With Modeling



In this scheme, advertisements are only given to customers who are predicted by the model to click on ads, namely 142 customers.

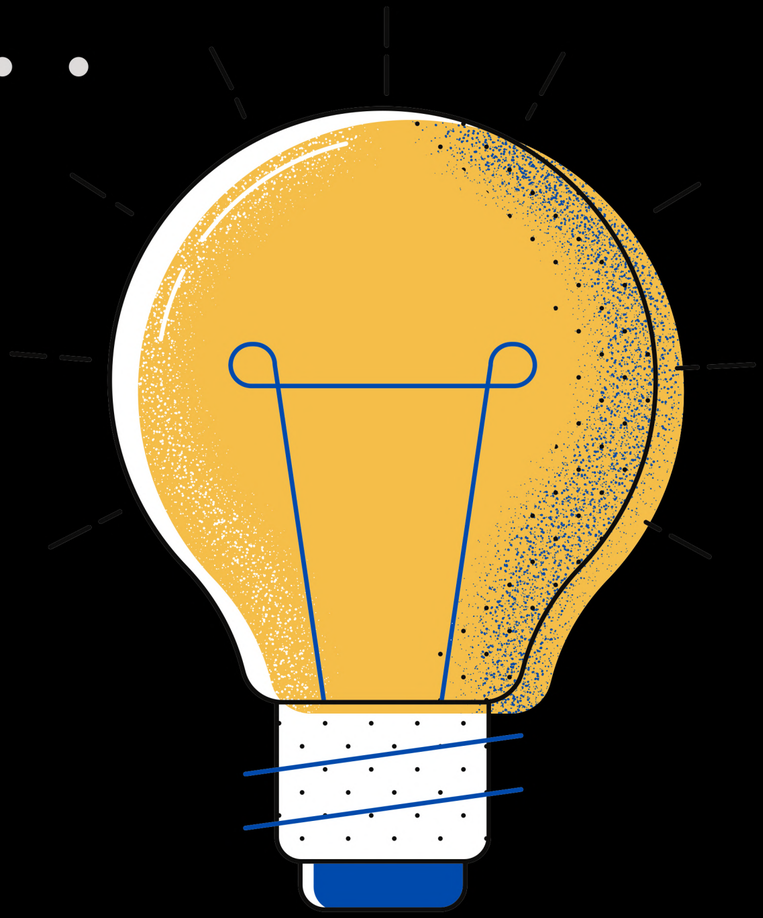
Business Simulation

The following is a comparison of the two business simulation schemes.

Without Modeling		With Modeling	
Click Through Rate (CTR)	= (144/298) * 100% = 48.32%	Click Through Rate (CTR)	= (140/142) * 100% = 98.59%
Total Cost (TC)	= 1000 * 298 = Rp. 298,000	Total Cost (TC)	= 142 * 298 = Rp. 142,000
Revenue (R)	= Rp. 5,000 * 144 = Rp. 720,000	Revenue (R)	= Rp. 5,000 * 142 = Rp. 700,000
Profit	= R – TC = Rp. 422,000	Profit	= R – TC = Rp. 558,000

- **By using machine learning CTR increased by 50.27%.**
- **Cost efficiency of 52.35%**, where in the business simulation it was found that the costs not incurred due to using the results from the model amounted to Rp.156,000.
- **The percentage of profit has increased by 32.32%** or from business simulation it is found that the difference in profit using the results of the model and not using the model is Rp. 136,000.

Thank You



Email
jonisyofian14@gmail.com

LinkedIn
https://www.linkedin.com/in/jonisyofian/