Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Azzam Majduddin
majduddin8@gmail.com
LinkedIn Profile

An intellectually curious and self-motivated statistics graduate with passion for technology seeking a meaningful role to begin a career in data. I have also completed a 4 month data science course organized by Rakamin Academy. Skilled at Data Processing, Data Visualization, Data Exploration, Programming languages, Microsoft Office, and Statistical Software.

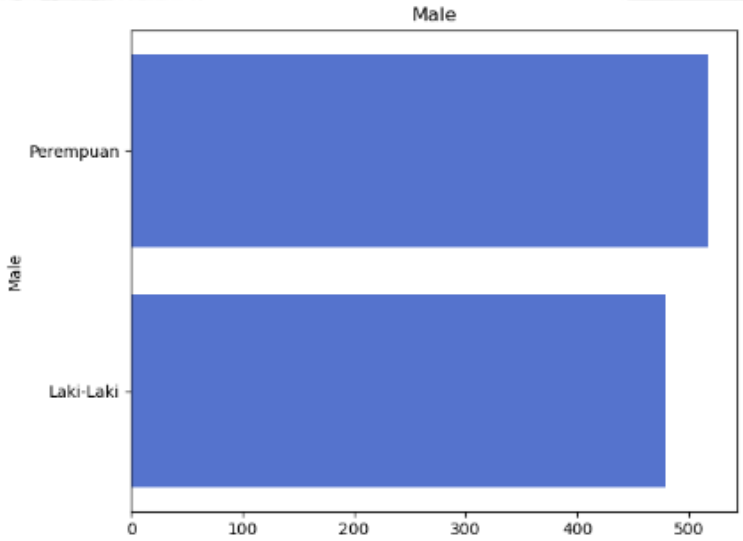
Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

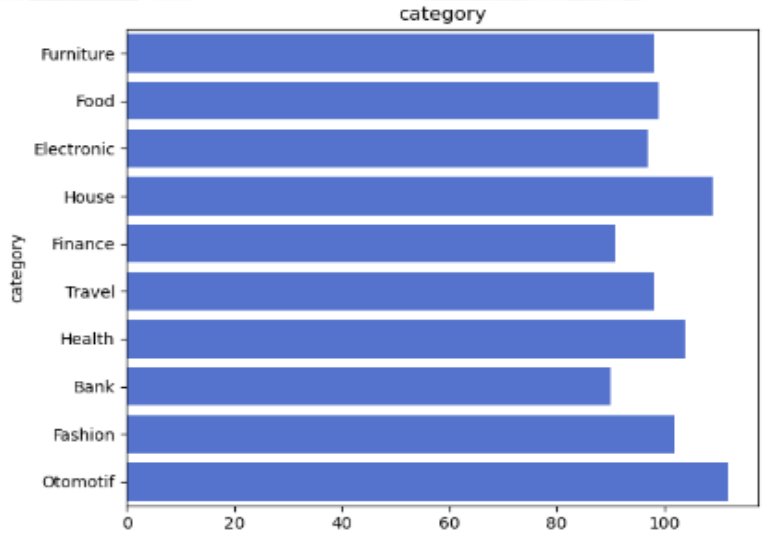
Overview

"A company in Indonesia wants to know the effectiveness of an advertisement that they are displaying, this is important for the company to find out how much the advertisement has reached in the market so that can attract customers to see the advertisement.

By processing the historical advertisement data and finding insight and patterns that occur, it can help companies determine marketing targets, the focus of this case is to create a machine learning classification model for determining target customers."

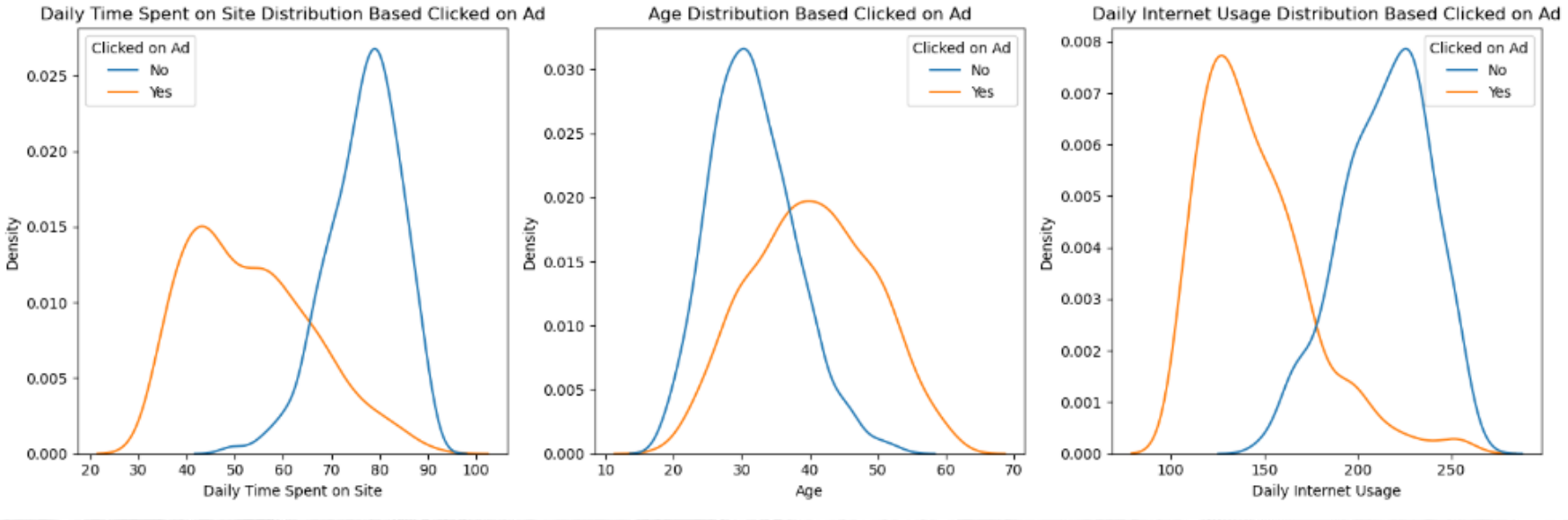
Customer Type and Behaviour Analysis on Advertisement





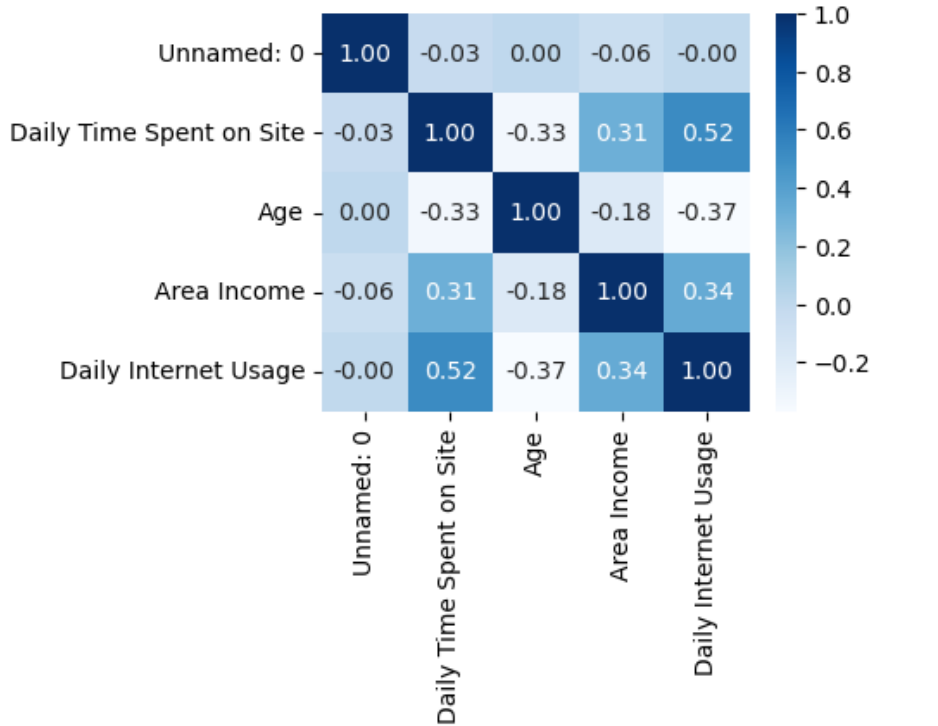
- Male (gender) feature dominated by female with 518 customers, while male are 479 customers.
- From category feature, Otomotif is the most clicked ads with 112 customers, followed by House with 109 customers.

Customer Type and Behaviour Analysis on Advertisement



- Customers with Daily Time Spent on Site around 40-45 minutes are clicking ads on the website, while customers who don't click on ads are with Daily Time Spent on Site around 75 - 80 minutes.
- Customers who aged 40 years old are the customers who click the most ads, while customers who aged 30 years old are the customers who don't click the ads the most.
- Customers with Daily Internet Usage around 100 – 150 tend to click on ads, while customers with Daily Internet Usage around 200 – 250 tend not to click on ads.

Customer Type and Behaviour Analysis on Advertisement



From the correlation heatmap beside, there is no multicollinearity (has a small correlation each column), so this features can be used for the modeling.

Data Cleaning & Preprocessing

Rakamin Academy

Handling Missing Values

There are 4 features that contain missing values. There are **Daily Time Spent on Site, Area Income, Daily Internet Usage**, and **Male (Gender)**. For the **Male** feature, missing value replaced by the mode of the data. For the other features, missing values replaced by the median of the data, because the other features are numeric data types.

Handling Missing Values ¶

```
: df.isnull().sum().sort_values(ascending = False)
```

```
: 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
  city                         0
  province                     0
  category                     0
  dtype: int64
```

Feature Extraction

```
# Converting timestamp column to datetime
df['Timestamp'] = pd.to_datetime(df['Timestamp'])
# Extracting Timestamp feature into Year, Month, Weekday, and Day
df['Year'] = df['Timestamp'].dt.year
df['Month'] = df['Timestamp'].dt.month
df['Weekday'] = df['Timestamp'].dt.dayofweek
df['Day'] = df['Timestamp'].dt.day
```

Feature Extraction

Timestamp features can be extracted into Year, Month, Weekday, and Day. So, the features are increase and can be used for the modelling.

Data Cleaning & Preprocessing

Rakamin Academy

Feature Encoding

```
# 'Male' feature
mapping_gender = {'Perempuan' : 0,
                  'Laki-Laki' : 1}
df['Male'] = df['Male'].map(mapping_gender)
```

```
# Variable Target
mapping_target = {'Yes' : 1,
                  'No' : 0}
df['Clicked on Ad'] = df['Clicked on Ad'].map(mapping_target)
```

```
# Category feature
df = pd.get_dummies(df, columns = ['category'])
```

Feature Encoding

Converting **Male, category, and Target/Clicked on Ad** features into numeric. For the **category** feature, feature encoding using one hot encoding; get_dummies().

Split Data Feature and Target

There are 19 features that used for the modelling, and for the target variable is **Clicked on Ad**.

Splitting Data into Feature and Target

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

Data Modeling

model	Recall_test	Recall_train	Accuracy_test	Accuracy_train	Time_Elapsed
Gradient Boost	0.948052	0.988439	0.943333	0.994286	15.62
XGBoost	0.941558	0.979769	0.940000	0.984286	25.16
Random Forest	0.941558	0.950867	0.943333	0.967143	58.44
AdaBoost	0.935065	0.988439	0.946667	0.994286	31.38
Decision Tree	0.837662	0.927746	0.893333	0.960000	14.02
Logistic Regression	0.824675	0.884393	0.893333	0.911429	21.55

Experiment 1

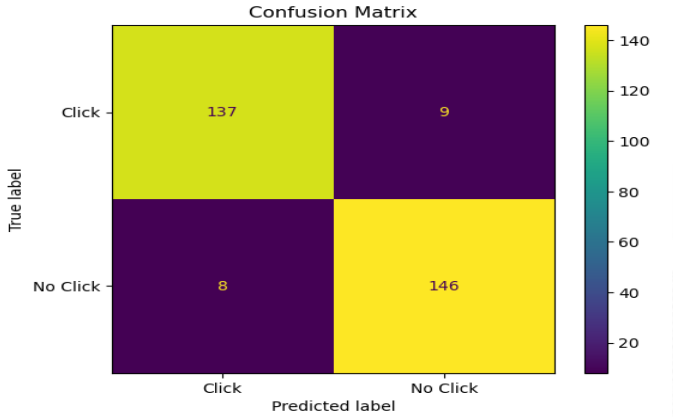
From the picture beside, can be concluded that Gradient Boost is the best algorithm because it has the highest recall score and accuracy score. XGBoost and AdaBoost also has a high accuracy and recall score, but Gradient Boost has a highest recall score than them.

Experiment 2

After converting the features into normalization, there is a significant increase in logistic regression model. Gradient Boost still the best model after converted the features into normalization.

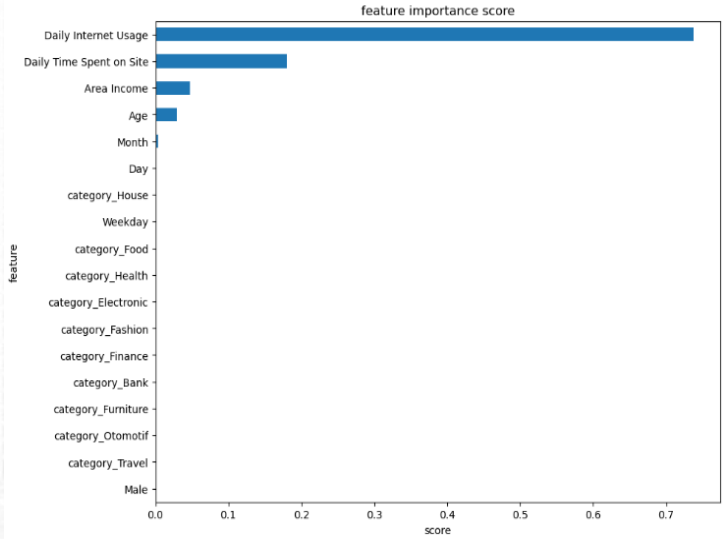
model	Recall_test	Recall_train	Accuracy_test	Accuracy_train	Time_Elapsed
Gradient Boost	0.948052	0.988439	0.943333	0.994286	0.00
XGBoost	0.941558	0.979769	0.940000	0.984286	15.62
Random Forest	0.941558	0.950867	0.943333	0.967143	82.86
AdaBoost	0.935065	0.988439	0.946667	0.994286	31.24
Logistic Regression	0.902597	0.927746	0.946667	0.962857	0.00
Decision Tree	0.837662	0.927746	0.893333	0.960000	14.32

Data Modeling

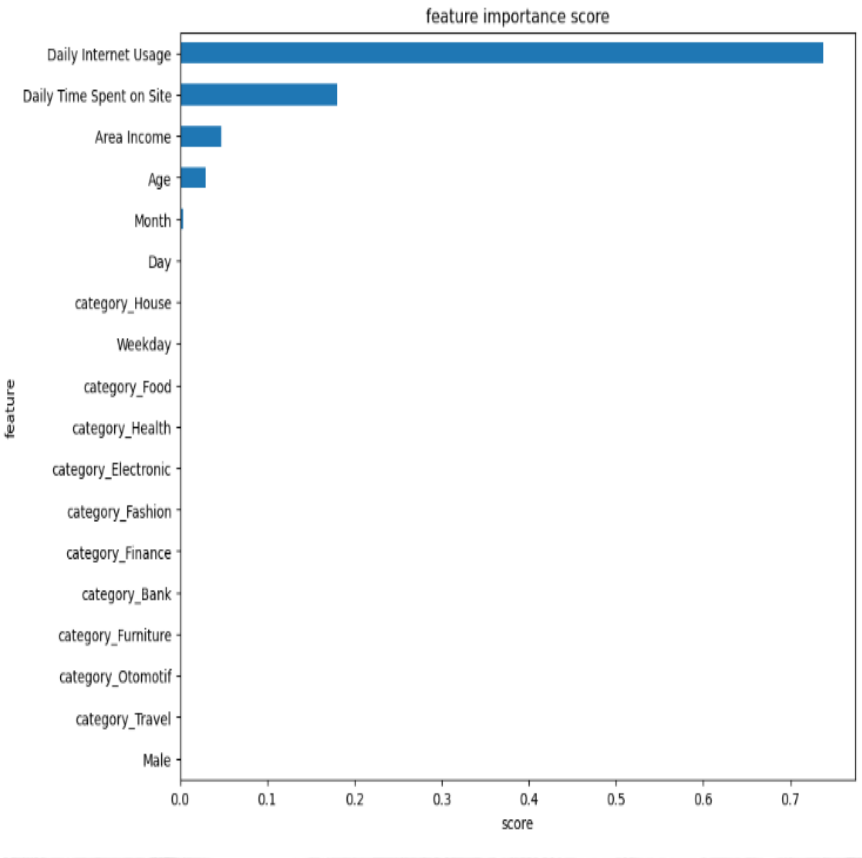


The confusion matrix from the gradient boost model is very good. The number of prediction errors is very small (8 False Positive and 9 False Negative). With the following result, we will get good accuracy score and recall score.

By using the gradient boost model, we can see which features are important in building the model. Two main features that can determine the success of marketing are **Daily Internet Usage** and **Daily Time Spent on Site**. The other features that affect the building of the model are **Area Income** and **Age**.



Business Recommendation & Simulation



Business Recommendation:

Based on EDA and Feature importance, can be concluded that:

- Daily Internet Usage is the most importance feature for building the model. From the Daily Internet Usage Distribution graph, the more the customers use the daily internet they are not interested in clicking on ads. While less daily internet uses, they are interesting in clicking on ads.
- Middle old is a potential advertising market, while the youths are not interested in clicking the ads. Action Needed: Creating a relevant advertisement for the youths such as fashion or educational supplies.

Simulation

Variable Target

```
y_test.value_counts()
```

```
1    154
0    146
Name: Clicked on Ad, dtype: int64
```

Before using machine learning model

Assumption:

- To advertise the customer, use a budget of Rp. 5,000
- Using the data testing as a simulation implementation with a total of 300 customer, 154 customer click ads and 146 customer didn't click ads.
- Every user who converts us will get a profit of Rp 7,500

Cost Calculation:

Cost = cost ads x n user

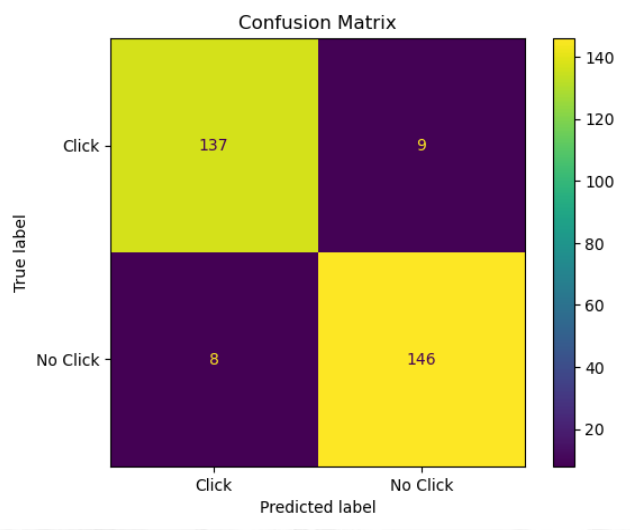
Cost = Rp. 5,000 x 300

Cost = Rp 1,500,000

- The conversion rate that we will get is 50%
- Because there are only 154 converts, we will get 154 x Rp 7,500 = Rp 1,155,000
- Profit = Revenue - Cost
  = Rp 1,155,000 – Rp 1,500,000
  = - Rp 345,000

Based on that simulation, we will loss Rp 345,000 if we didn't use the machine learning model.

Business Recommendation & Simulation



After using machine learning model

- Cost 145 x Rp 5,000 = 725,000
- 137 customers are convert
- Revenue = 137 x Rp 7,500 = 1,027,500
- Profit = Revenue - Cost
  = Rp 1,027,000 – Rp 725,000
  = Rp 345,000

Based on the simulation, we will get Rp 345,000 profit. In conclusion, machine learning work properly to increase the profit rather than without using machine learning model that we will get loss.