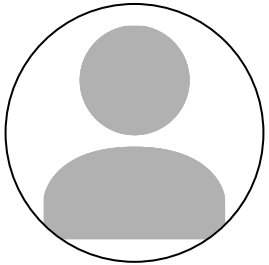
Predict Clicked Ads Customer Classification by using Machine Learning

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



Created by:
Tsabitah Karimah
tsabitahkarimah@gmail.com
LinkedIn

"I am a dietetics graduate with a newfound interest for data science, seeking to leverage analytical skills and healthcare knowledge in a data scientist role. Completed an intensive 5-month data science course, gaining proficiency in key programming languages and machine learning techniques. Eager to apply newly acquired skills in data manipulation, visualization, and predictive modeling to extract meaningful insights from complex datasets. Combines technical acumen with strong problem-solving abilities to drive data-informed decision-making in a data scientist role."

Overview

"An Indonesian company wants to evaluate the effectiveness of an advertisement they've aired. This is crucial for the company to understand the extent to which their advertisement has reached its target audience and successfully attracted customers.

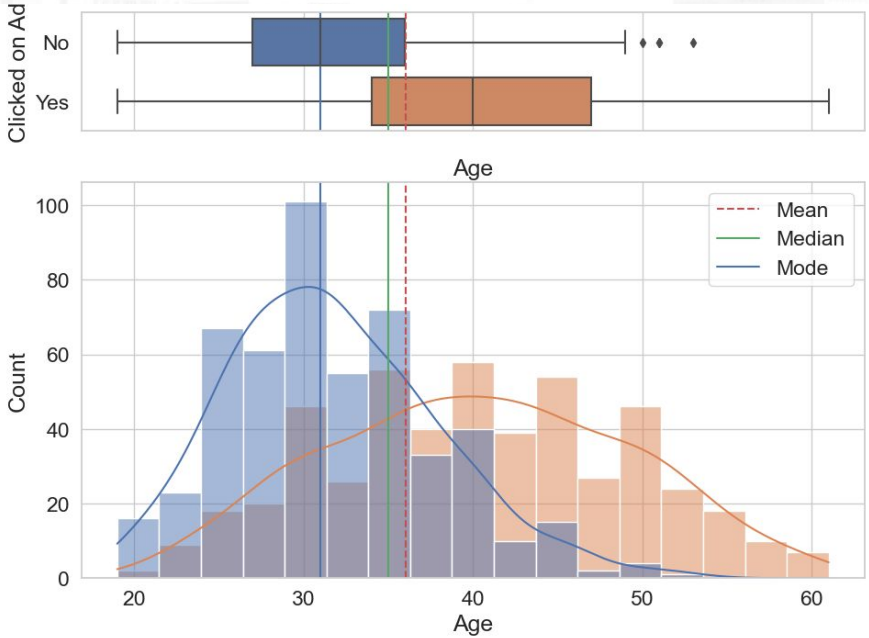
By processing historical advertisement data and uncovering insights and patterns, the company can better determine its target market. The focus of this case is to create a machine learning classification model that can accurately identify the right target customers."

Exploration Data Analysis (EDA)

- At a glance, the data has 1000 rows and 11 columns
- There are 4 columns with missing values;
  - ‘Daily Time Spent on Site’
  - ‘Daily Internet Usage’
  - ‘Area Income’
  - ‘Male’
- All data values matched with the data type except for timestamp
  - Hence, the data type was converted from object to timestamp
- All the features then grouped based on its data type.

For further details, please refer to this Jupyter Notebook

Univariate analysis - Age

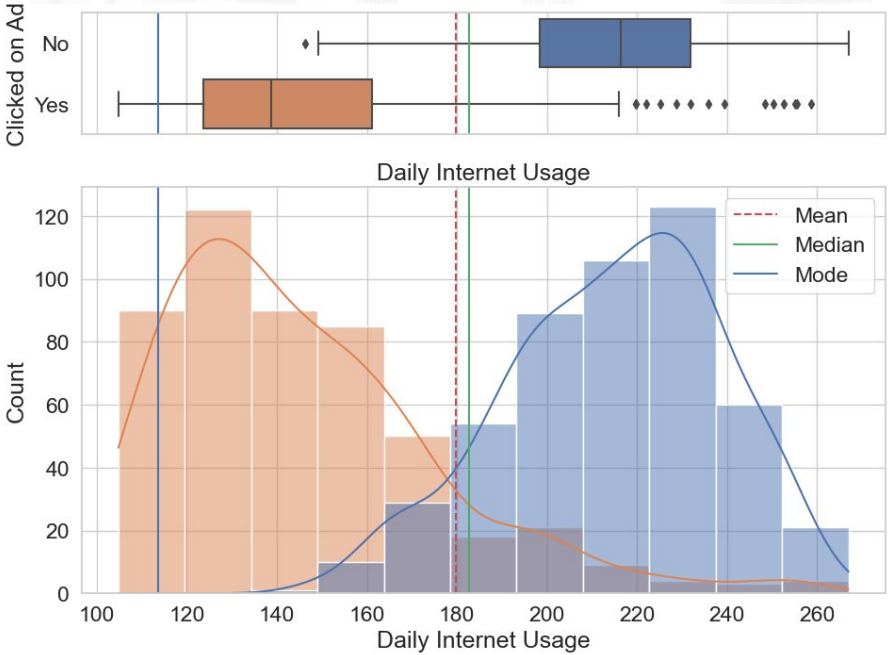


- Customer's age is ranging from 19-61 years old
- The data is slightly positive skewed as the mean (36 years old) is higher than the median (35 years old)
- The median age of those who clicked on the ad appears to be slightly lower than the median age of those who didn't.
- The histogram suggests that the age distribution for those who didn't clicked on the ad might be slightly skewed towards younger ages compared to those who did.
- The box plot shows a number of outliers in the "No" group, indicating some individuals in this group are significantly older or younger than the majority.

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Univariate analysis - Daily Internet Usage

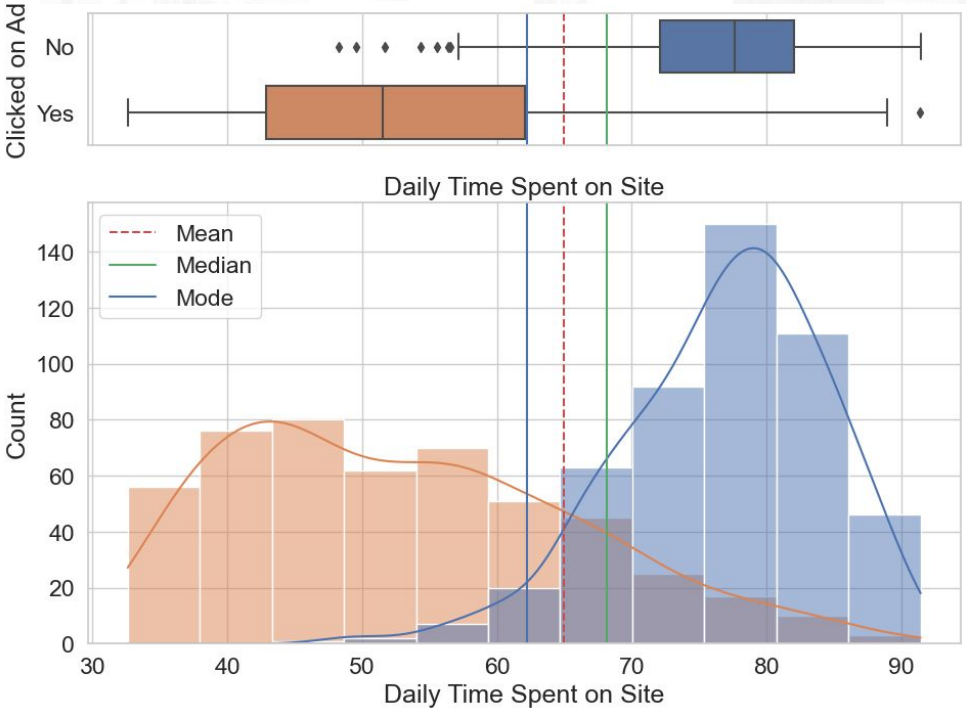


- Daily internet usage among customers is ranging from 105-267 mb
- The data distribution is slightly negative skewed as shown by the median (183mb) higher than the mean (180mb)
- The median daily internet usage of those who clicked on the ad appears to be slightly higher than the median usage of those who didn't.
- The box plot shows a few outliers in both groups, indicating some individuals have significantly higher or lower daily internet usage.
- The histogram suggests that the distribution of daily internet usage for those who didn’t clicked on the ad might be slightly skewed towards higher usage compared to those who did.

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

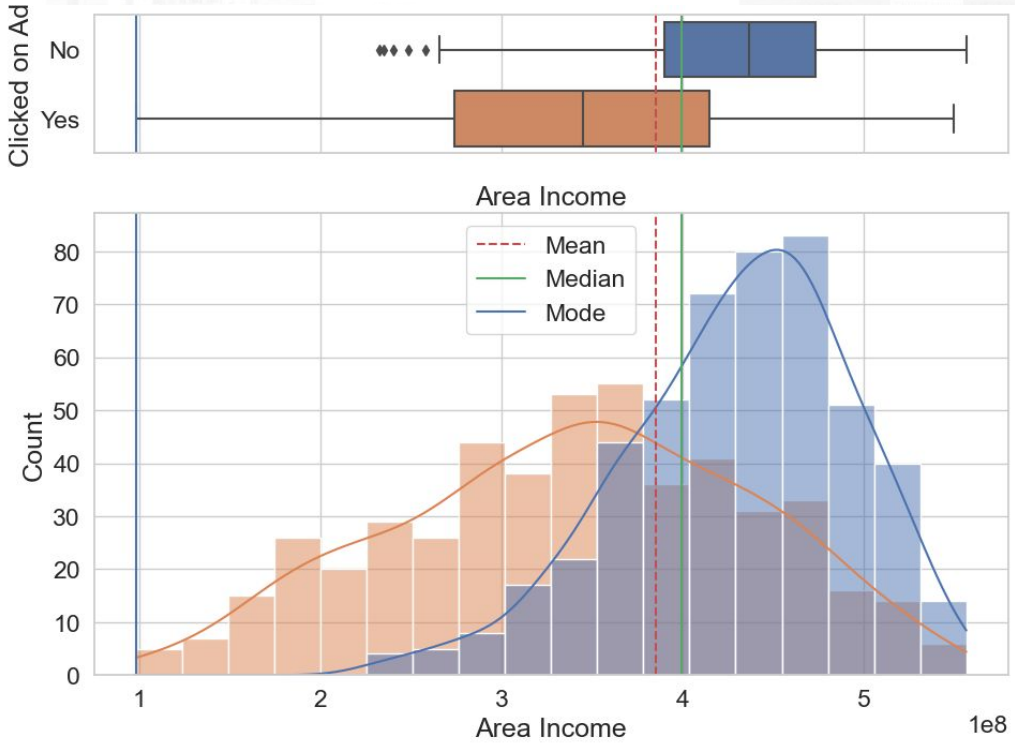
Univariate analysis - Daily Time Spent on Site



- Daily time spent on site ranging from approximately 33 mins to 91 mins
- The data distribution is slightly negative skewed as shown by the median (68 mins) that is higher than the mean (65 mins)
- The box plot shows numbers of outliers of those who didn't clicked on the ad.
- The median daily time spent on the site for those who clicked on the ad appears to be slightly higher than the median for those who didn't.
- The histogram suggests that the distribution of daily time spent on the site for those who did not clicked on the ad might be slightly skewed towards higher usage compared to those who did.

For further details, please refer to this Jupyter Notebook

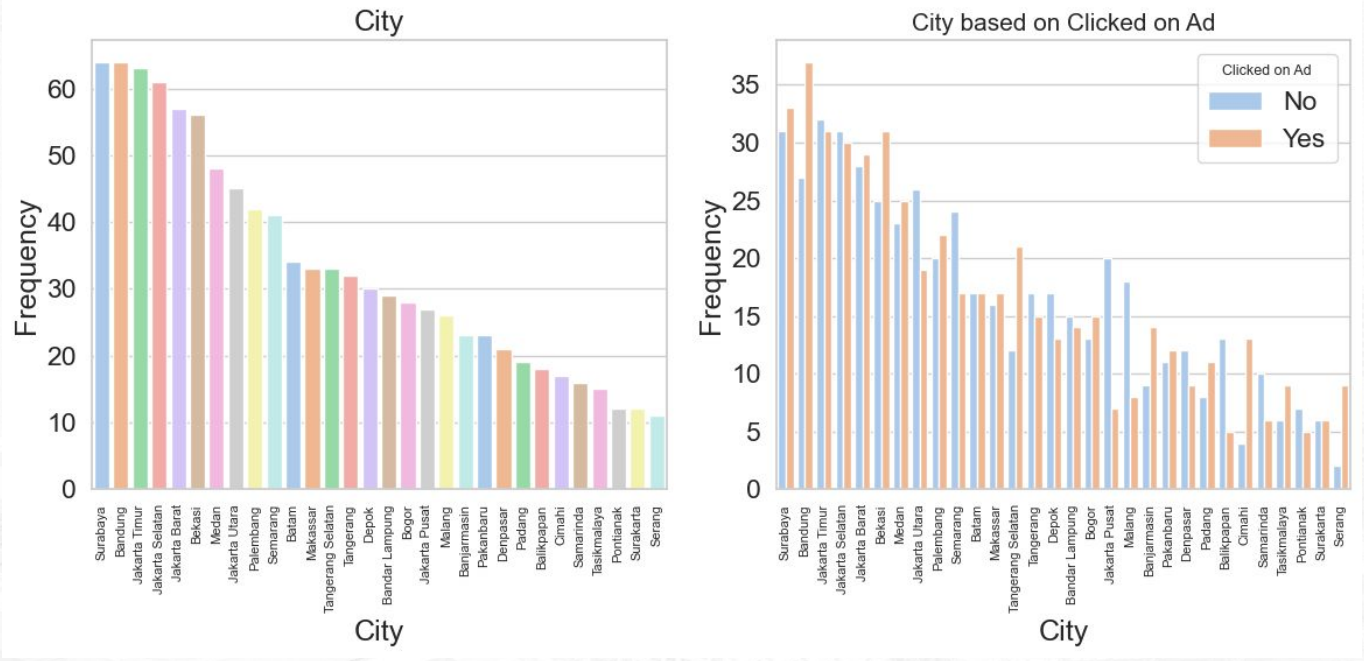
Univariate analysis - Area Income



Area Income: in the range of 97.975.500 - 556.393.600 million, median (399.068.320) > mean (384.864.670), slightly negative distribution

For further details, please refer to this Jupyter Notebook

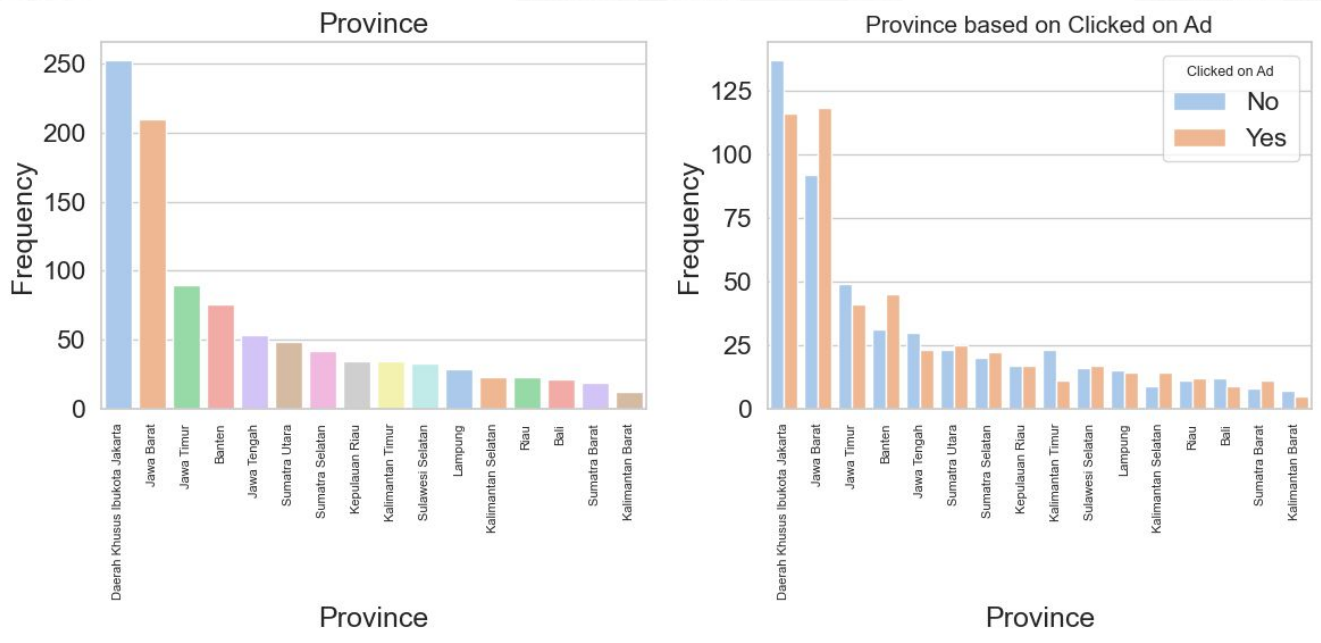
Univariate analysis - City



- The customers are located in a total of 30 cities with Surabaya and Bandung as the city where most customers are located.
- Customers who clicked on ad mostly located in Bandung, Surabaya, and Bekasi.

For further details, please refer to this Jupyter Notebook

Univariate analysis - Province

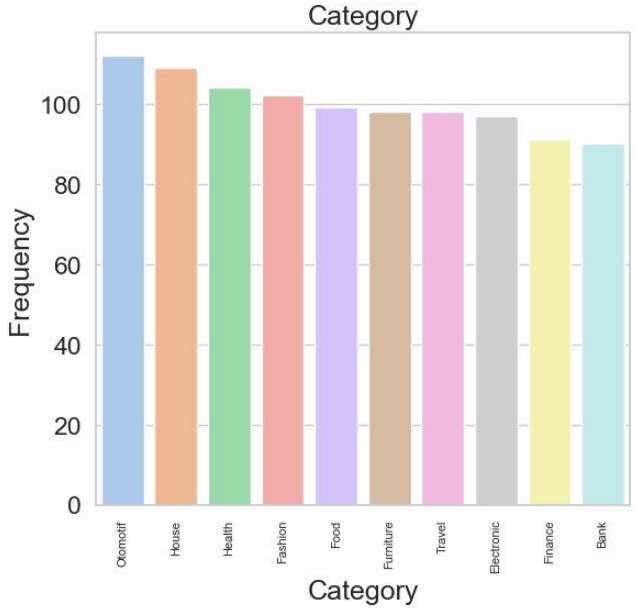


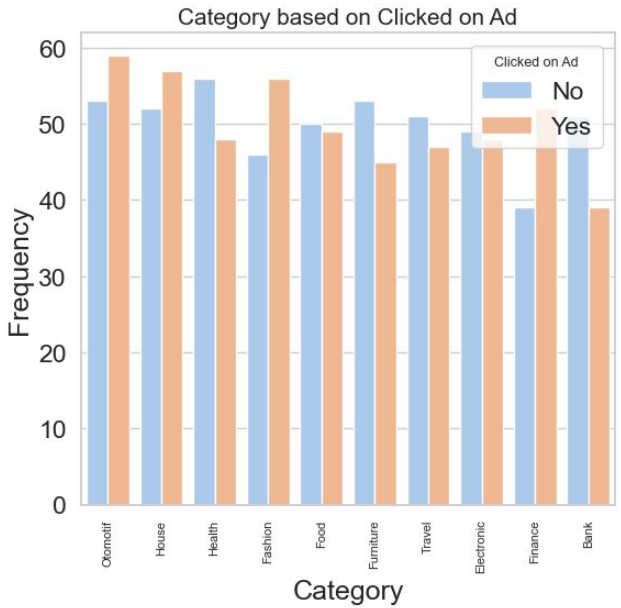
- Customers are located in 16 province with most customers who clicked on the ad located in DKI Jakarta and Jawa Barat.

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Univariate analysis - Category



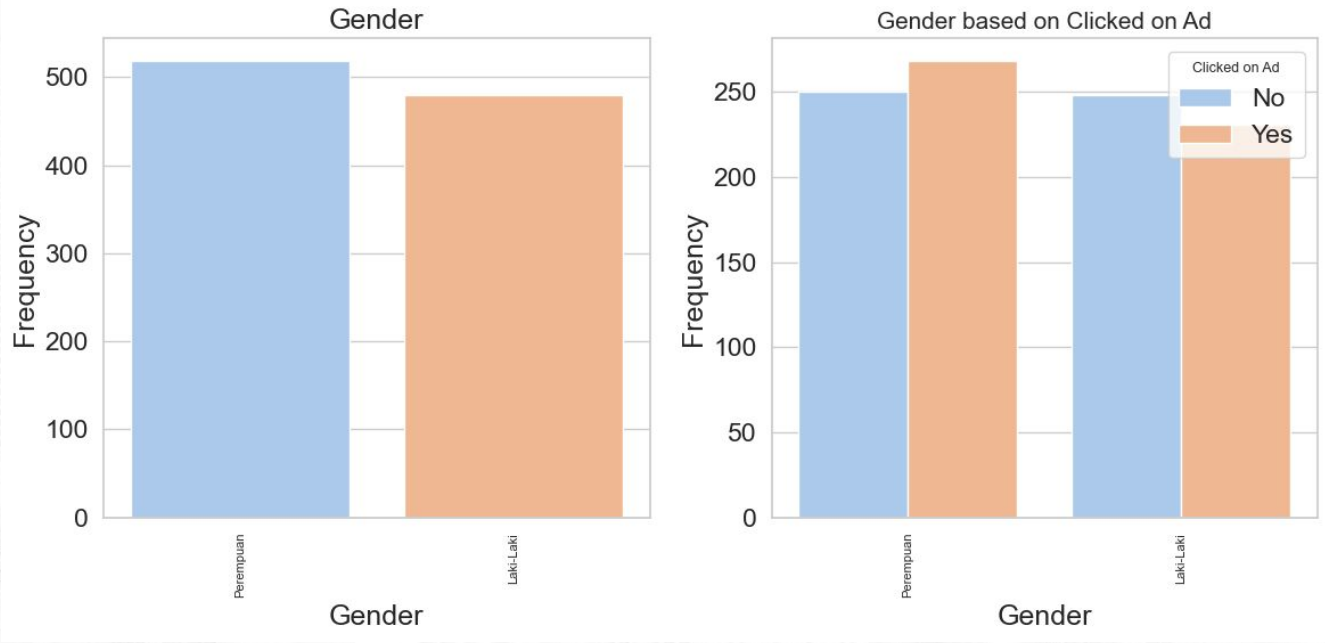


- The category most selected by the customers is Otomotif, followed by House, Health and Fashion.
- The least selected categories are Electronic, Finance, and Bank.
- Customers who clicked on the ad are mostly from category Otomotif, House, and Fashion.

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Univariate analysis - Male

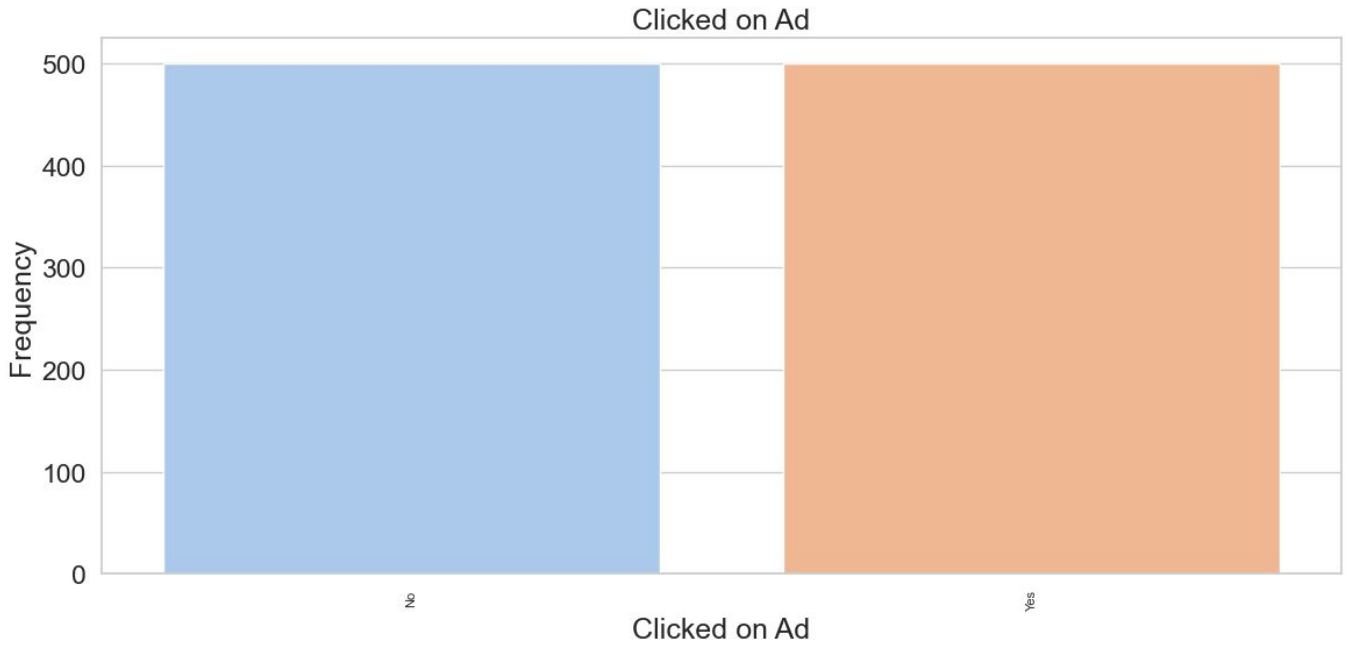


- Most of the customers are female and so are those who clicked on the ad

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

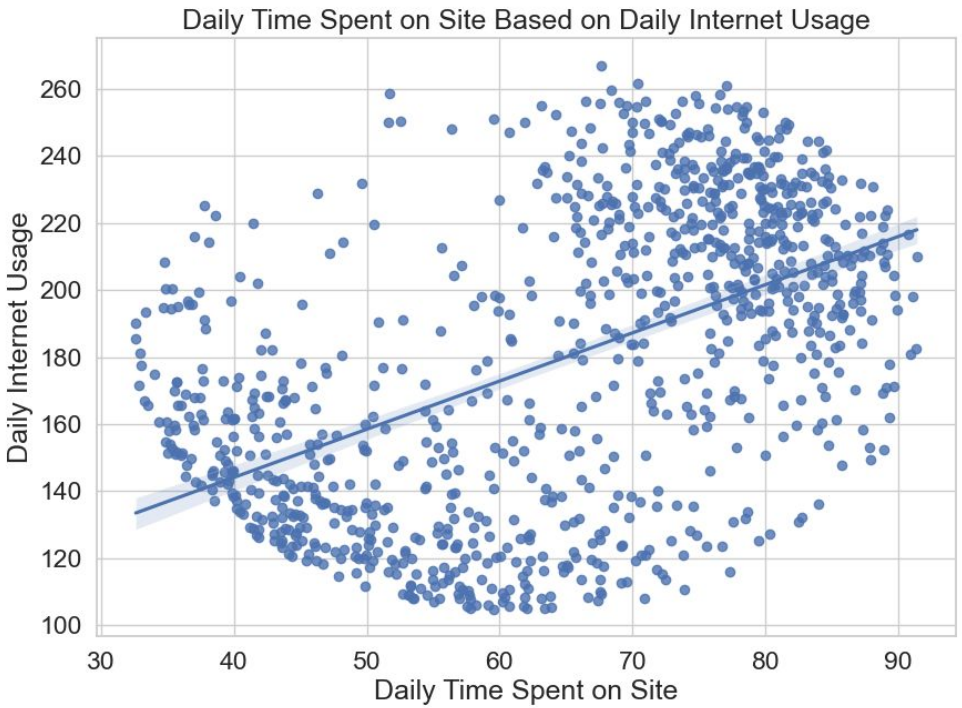
Univariate analysis - Clicked on Ad



- The number of customers who clicked and didn't click the ad are equal

For further details, please refer to this Jupyter Notebook

Bivariate analysis

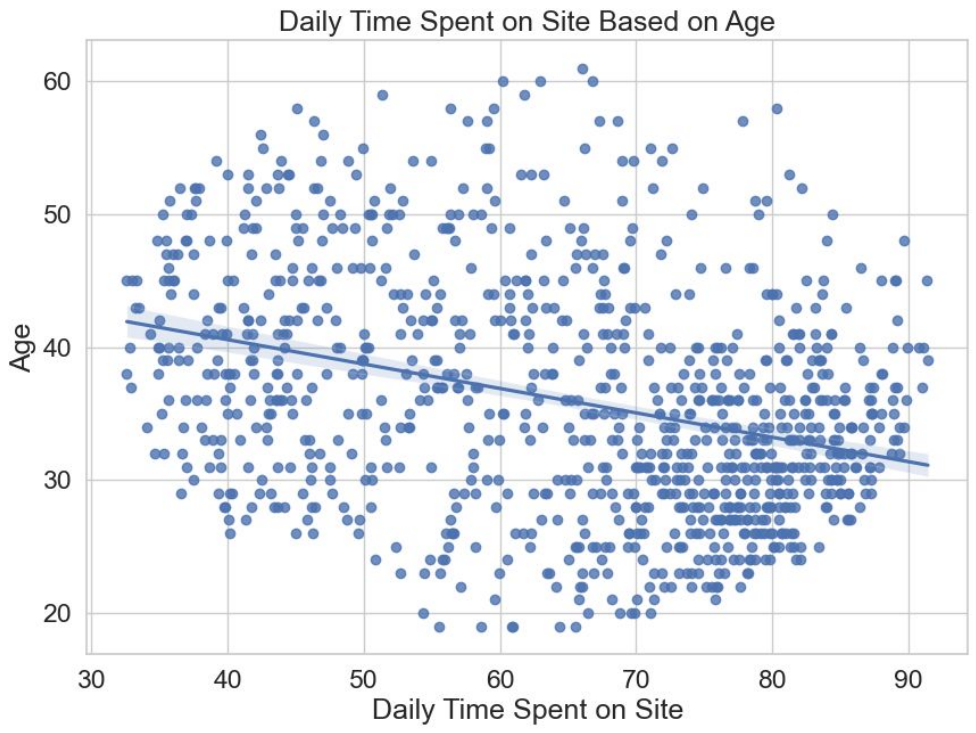


- The scatter plot shows positive relationship between "Daily Time Spent on Site" and "Daily Internet Usage"
- This implies that as the customer spent more time on the site, the daily internet usage also increases.

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Bivariate analysis

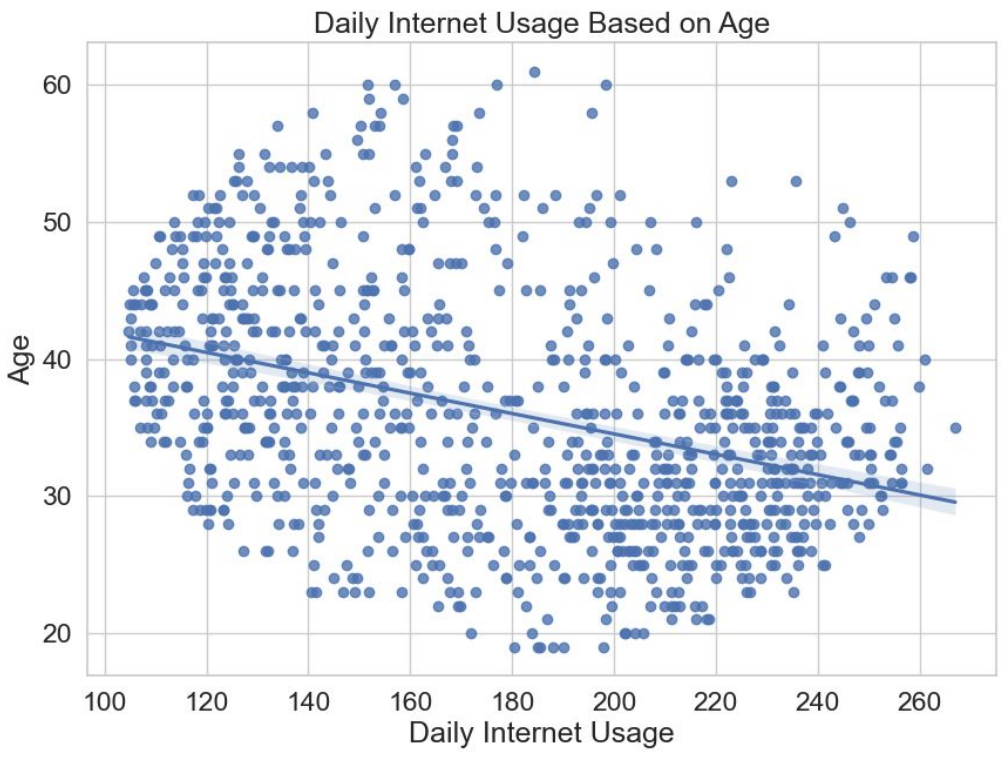


- The scatter plot shows negative relationship between "Daily Time Spent on Site" and "Age"
- This implies that the older the customer, the daily time spent on site is lesser

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Bivariate analysis



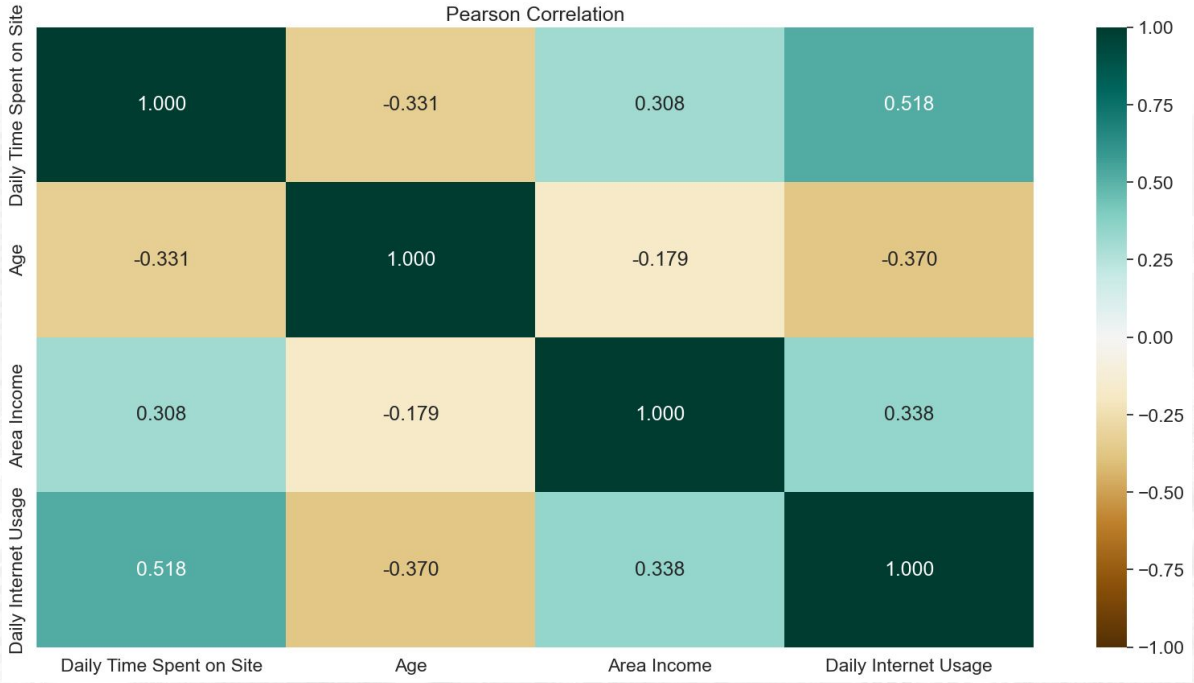
- The scatter plot shows positive relationship between "Daily Internet Usage" and “Age”
- This implies that as the age declines, the lesser daily internet usage

For further details, please refer to this Jupyter Notebook

Customer Type and Behaviour Analysis on Advertisement

Rakamin Academy

Multivariate analysis



- Strong positive correlation can be seen between “Daily Time Spent on Site” and “Daily Internet Usage” (r=0.518).
- Moderate positive correlation can be seen between “Daily Time Spent on Site” and “Area Income” (r=0.308). Customers with higher income tend to spend more time on site.
- Weak negative correlation can be seen between “Daily Time Spent on Site” and “Age” (r=-0.370)
- There is no significant association between “Age” and “Area Income” (r=-0.179).

For further details, please refer to this Jupyter Notebook

Missing Values & Duplicated Value

```
Daily Time Spent on Site     13
Age                           0
Area Income                  13
Daily Internet Usage         11
Male                          3
Timestamp                     0
Clicked on Ad                 0
city                          0
province                      0
category                      0
```

- The above shows 4 columns with missing values which are **'Daily Time Spent on Site'**, **'Area Income'**, **'Daily Internet Usage'** and **'Male'** along with the total number of missing values for each column
- **Median** of each numerical column is used to fill in missing values
- Meanwhile, **mode** of categorical column (‘Male’) is used to fill in missing values
- There is **no duplicate value** in this data set

For further details, please refer to this Jupyter Notebook

Extract datetime data

Four new columns are extracted from column ‘Timestamp’

- Column 'Date_Day' shows the day represented by 0 as Monday, 1 as Tuesday, 2 as Wednesday, 3 as Thursday, 4 as Friday, 5 as Saturday, and 6 as Sunday
- Column 'Date_Month' shows the month based on the given timestamp represented by numerical value
- Column 'Date_Year’ shows the year in numerical value from the given timestamp
- Column 'Date_Week' shows the week in a year from the given timestamp
- Column ‘Timestamp’ then dropped s it has represented by the extracted data

For further details, please refer to this Jupyter Notebook

Feature Encoding

- **Ordinal Encoding** is done for column **‘Clicked_on_Ad’** and **‘Male’**
- **One Hot Encoding** is done for column **‘city’**, **‘province’** and **‘category’**
- These 5 columns are then **dropped** as it has represented by the encoded columns

For further details, please refer to this Jupyter Notebook

Feature Selection

Feature selection is done to drop irrelevant and redundant columns

- Drop features that has **high correlation r>0.70**

- Encoded **province columns that has high correlation with encoded city columns are dropped** and the city columns are retained as city has more detailed information

For further details, please refer to this Jupyter Notebook

Split data

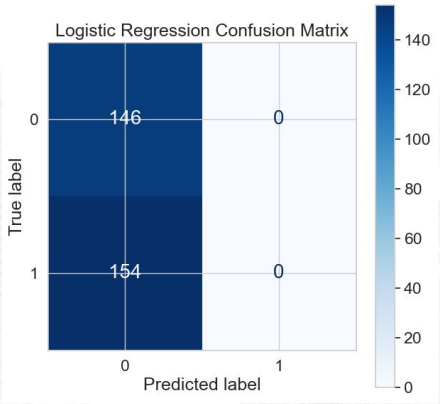
- X represents the features data
  - This includes all the data after cleaning and preprocessing is done except for column ‘clicked_on_ad’
  - It has a total of 51 columns

- y represents the target data from column ‘clicked_on_ad’

For further details, please refer to this Jupyter Notebook

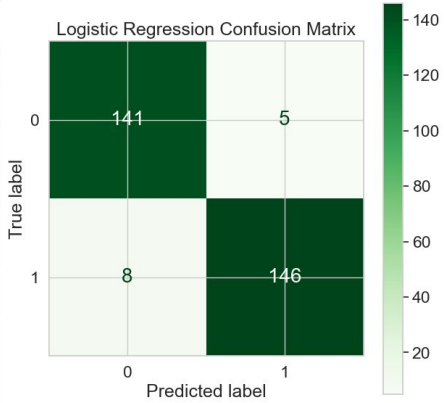
Logistic Regression

1st Experiment Result




```
'Accuracy': 0.49,
'Precision': 0.00,
'Recall': 0.00,
'F1-Score': 0.00
```

2nd Experiment Result



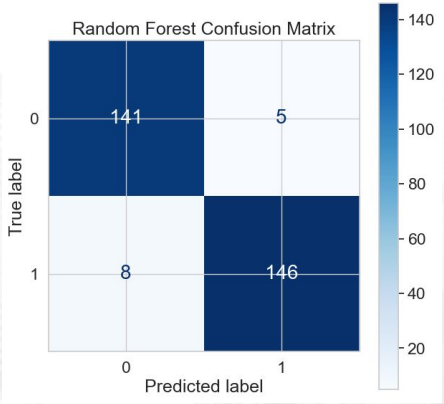

```
'Accuracy': 0.96,
'Precision': 0.97,
'Recall': 0.95,
'F1-Score': 0.96
```

- Standardization has significantly improved the model's accuracy, precision, recall, and F1-score
- Prior to standardization, the model has low accuracy (49%) and it significantly improves to 95%
- This experiment shows the importance of data preprocessing, especially standardization to improve the performance of machine learning models.

For further details, please refer to this Jupyter Notebook

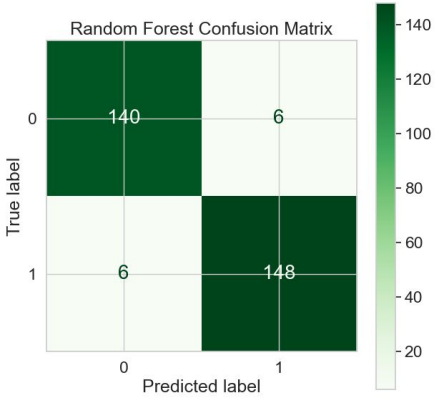
Random Forest

1st Experiment Result




```
'Accuracy': 0.96
'Precision': 0.97
'Recall': 0.95
'F1-Score': 0.96
```

2nd Experiment Result (Standardized)

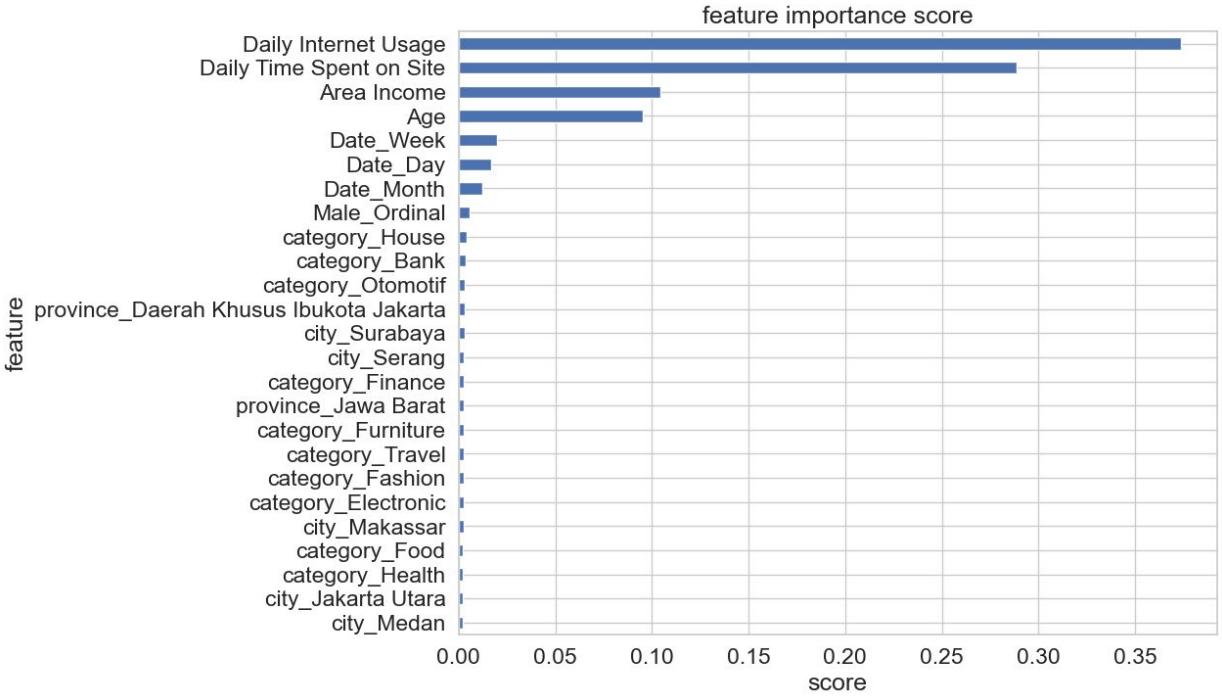



```
'Accuracy': 0.96
'Precision': 0.96
'Recall': 0.96
'F1-Score': 0.96
```

- Generally, all the scores are similar before and after standardization. Both shows high accuracy, precision and recall and F1 score.
- In contrast of Logistic Regression, this model shows a good performance before and after standardization, showing its robustness to standardization.

For further details, please refer to this Jupyter Notebook

Feature Importance (Random Forest)



- The figure shows the hierarchy of the most important features to click ad classification
- Four top features are similar to the previous figure
- Week of the year and day are the next most important features that affect click ad classification

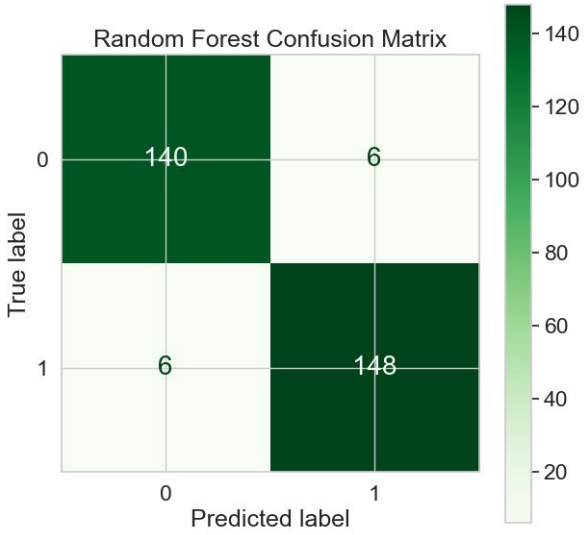
For further details, please refer to this Jupyter Notebook

Business Recommendation based on EDA and Feature Importance

- **Target users with high engagement**
  - Target users with higher usage levels (above 180 Mb) as they are more likely to be engaged online.
  - Target users with time spent more than 1 hour
- **Use age-targeted advertising**
  - Creating ads to specific age groups especially to younger age groups, focusing on products or services that are relevant to their interests and lifestyle
- **Target higher income users**
  - for products and services that are more expensive
- **Further analysis on specific day, week and month where customers are more likely to click on the ad**

For further details, please refer to this Jupyter Notebook

Business Recommendation & Simulation



Simulation

- **Before using Machine Learning (Number of Users: 300)**
  - Number of customers: 300
  - Cost: 5.000/customer = 1.500.000
  - Conversion rate: 50%
  - Average purchase value: 10.000
  - Only 150 actually click the ad
    - Revenue: 1.500.000
    - Profit: 0
- **After using Machine Learning (Number of Users: 300)**
  - Predicted Clicked customers: 154
  - Cost: 5.000/customer = 770.000
  - Conversion rate: 96.1%
  - Average purchase value: 10.000
  - Only 148 actually click the ad
    - Revenue: 1.480.000
    - Profit: 710.000

For further details, please refer to this Jupyter Notebook

Simulation

- Before implementing a machine learning model, this company conducted marketing randomly, such as sending promotional emails to all customers without clear segmentation.
  - Costs of creating marketing content (graphic design, copywriting), and online advertising costs (e.g., Google Ads)
  - Revenue from product sales
  - Profit = Revenue - Costs

- After Using a Machine Learning Model, the company can perform better customer segmentation. For example, the model can identify customers who are more likely to purchase new products based on their purchase history, browsing behavior, and demographics.
  - Costs of developing and implementing the machine learning model, maintenance costs for the model, and costs of more targeted marketing campaigns (e.g., Facebook Ads).
  - Revenue: Increased revenue due to more effective marketing campaigns and higher conversion rates as products are more relevant to customer interests.
  - Profit: Increased profit due to a reduction in ineffective marketing costs and increased revenue.

- Overall, the simulation shows that using machine learning models in marketing can provide numerous benefits, including increased cost efficiency, higher revenue, and improved customer satisfaction

For further details, please refer to this Jupyter Notebook