Suryansh Mishra

Predict Clicked Ads Customer Classification

Tool : Jupyter Notebook
Programming Language : Python
Libraries : Pandas, NumPy, Scikit Learn, XgBoost
Visualization : Matplotlib, Seaborn

📂 Introduction

Background

As times change, companies must optimize their advertising methods on digital platforms to attract potential customers at minimal cost. The goal is to increase conversions, i.e., the number of potential customers who make a purchase after clicking on an ad. However, to achieve this, companies need to be able to accurately predict click-through rates. Accurate click-through rate predictions are crucial for determining the success of digital advertising campaigns. Without accurate predictions, companies may incur high costs without significant results.

Goal

To create a machine learning model that can identify potential users likely to convert or be interested in an ad, enabling the company to optimize its advertising costs.

Objective

- Predict users who are likely to click on ads with 90% accuracy.
- Gain insights into potential user patterns that lead to ad clicks.
- Provide business recommendations based on analysis and model results.

📂 STAGE 1: Exploratory Data Analysis

Data Overview

The dataset has 1000 rows and 9 features, along with 1 target variable. Here is a description of the dataset features:

Feature	Description
Daily Time Spent on Site	Daily time spent on a site (in minutes)
Age	User's age in years
Area Income	User's income (in currency)
Daily Internet Usage	Daily internet usage (in minutes)
Male	User’s gender
Timestamp	When the user visited a site
Clicked on Ad	Whether or not the user clicked on an ad
City	User's city
Province	User's province
Category	Product category visited

Data Quality Assessment

Data assessment is performed to ensure that the data is ready and suitable for further analysis. The following checks were conducted:

- **Checking for missing values:** Missing values found in features like Daily Time Spent on Site, Area Income, Daily Internet Usage, and Male. Missing values in numerical features were addressed using median imputation, while categorical features, such as Age, used mode imputation.
- **Checking for duplicates:** No duplicates were found.
- **Feature or value consistency:** An unused feature (Unnamed: 0) was dropped, and the Timestamp data type was converted to datetime format, with month, week, day, and hour extracted.

- **Outliers or anomalies:** Outliers were found in Area Income, but they were tolerable as they were not extreme.

Data Assessment	Finding	Handling
Missing values	Present in Daily Time Spent, Area Income, Daily Internet Usage, Male	Handled by median imputation for numerical data, mode for categorical data
Duplicates	None	No handling required
Inconsistent features	- Unused feature: Unnamed: 0 - Incorrect data type: Timestamp	Dropped Unnamed: 0, converted Timestamp to datetime
Outliers	Area Income contains outliers	They were handled

Data Exploration

Customer Type and Behaviour Analysis on Advertisement

This analysis helps understand customer profiles and behaviours related to ads. Features used for this analysis include Daily Internet Usage, Daily Time Spent, and Age.

- **Daily Internet Usage** gives insights into how engaged customers are with online activities, helping identify customer groups likely to be more active online.
- **Daily Time Spent** reflects the time customers spend daily on online activities, indicating their engagement with digital content and potential to view or interact with ads.
- **Age** provides clues about the preferences of different customer age groups, as younger generations might be more open to technological innovation, while older generations might prefer content relevant to daily life.

Plot Analysis

- **Daily Time Spent Analysis** found that users spending less than 1 hour on a site have a higher potential to click on ads. This might be because they are more open to exploring ads, making them more likely to click.
- **Daily Internet Usage Analysis** showed that users who use the internet less frequently have a higher potential to click on ads, possibly due to curiosity about the products or services advertised.

- **Age Analysis** indicated that older users are more likely to click on ads. Younger users, who are more accustomed to the internet, may prefer other sources of information and tend to avoid intrusive ads.

Correlation Analysis

- **Daily Time Spent on Site vs. Internet Usage:** The correlation plot revealed two segments: active and non-active users. Active users spend more time on sites but tend to avoid clicking on ads. Based on this finding, companies could focus ad targeting on non-active users, who might be more responsive to ads. Adjusting content to appeal to these users could enhance ad campaign effectiveness.

📂 STAGE 2: Data Pre-processing

The following preprocessing steps were undertaken:

Handling Null Values, Check Duplicated Data, Feature Encoding (Clicked on Ads), Handling Outliers.

Feature Processing	**Description**
Selected Features	Daily Internet Usage, Daily Time Spent, Age,

---

📂 STAGE 3: Data Modelling and Evaluation

Model Experiment

Two different experiments were conducted for ad-click prediction. First used default training data, and the second applied StandardScaler standardization to the data due to its approximately normal distribution.

Observation without standardization:

- Tree-based models have far better performance than distance-based models.

- The best performance models are Random Forest, Extra Trees , Decision Tree, with the highest accuracy.
- The worst performance models are KNNeigbors and Logistic Regression with the lowest accuracy.
- The longest time elapsed occurred on Random Forest, Extra Trees , logistic regression and XGBoost models.

Observation with Standardization:

- After the dataset was normalized,The best performance model was the same as Random Forest
- The worst performance model still Logistic Regression although its performance increase significantly.
- The longest time elapsed still occurred on Random Forest, Extra Trees and XGBoost models

---

📂 STAGE 4: Business Recommendation & Simulation:

Business Recommendation

Based on the insight from EDA and feature importances, we can provide business recommendations such as:

Content Optimization

Because the higher Daily Time Spent on Site and Daily Internet Usage the less likely user will click on ads, then we need create ad contents that are engaging and relevant to the target user and ensure that the messaging and visuals of the ads align with the interests and needs of the user.

Targeted Pricing Strategies

Because the lower Area Income the more likely user will click on ads, we can implement targeted pricing strategies that align with the income levels of the target audience. This may involve creating special pricing tiers, discounts, or bundled offerings. Consider developing and promoting affordable products or services for the users with low area income.

Age-Targeted Marketing Campaigns

Because the older the user the more likely user will click on ads, then we can develop targeted marketing campaigns specifically designed to resonate with older demographics. We can create the messages, visuals, and offers to align with the preferences and interests of older users.

Business Simulation:

Assumption:

Cost per Mille (CPM) = Rp.100,000

Revenue per Ad Clicked = Rp.2,000

Before Using Machine Learning Model:

Number of Users Advertised:

User = 1,000

Click-Through Rate (CTR):

500/1,000 = 0.5

Total Cost:

CPM = Rp.100,000

Total Revenue:

CTR x Number of Users Advertised x Revenue per Ad Clicked = 0.5 x 1,000 x 2,000 = Rp.1,000,000

Total Profit:

Total Revenue - Total Cost = Rp.900,000

After Using Machine Learning Model:

Number of Users Advertised:

User = 1,000

Click-Through Rate (CTR):

Precision = 0.95

Total Cost:

CPM = Rp.100,000

Total Revenue:

CTR x Number of Users Advertised x Revenue per Ad Clicked = 0.95 x 1,000 x 2,000 = Rp.1,900,000

Total Profit:

Total Revenue - Total Cost = Rp.1,800,000

Conclusion:

From the results above, it can be seen that after we used the machine learning model, the ad performance increased. Click-Through Rate (CTR) increased 45% from 50% to 95% and total profit increased 100% from Rp.900,000 to Rp.1,800,000.

Sessions and Click on Ads

Analyzing the potential time for users who click on ads is important because it can provide valuable insights into user behavior and help companies optimize their marketing strategies.