Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Saila Fikriyya
sailafs@gmail.com
linkedin.com/in/sailafs1203/

I'm a health psychology graduate turned data scientist, bridging the gap between human behavior and data-driven solutions. With a Master's in Health Psychology from the University of Stirling and ongoing training in data science, I combine expertise in research methods, data analysis, and understanding human behavior. My experience spans from studying post-traumatic growth in cancer survivors to applying data science techniques in digital healthcare and career development. I'm passionate about leveraging interdisciplinary skills to create evidence-based strategies that improve health outcomes and overall well-being, always prioritizing a human-centric approach in collaborative environments.

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

Overview

Rakamin Academy

"A company in Indonesia wants to know the effectiveness of an advertisement they launched online. This is important for the company to understand how well the advertisement reaches its target audience, so they can attract customers to click on the ad.

By processing historical advertisement data and uncovering insights and patterns, it can assist companies in determining their marketing targets. The focus of this case is to create a machine learning classification model that serves to identify the right target customers. "

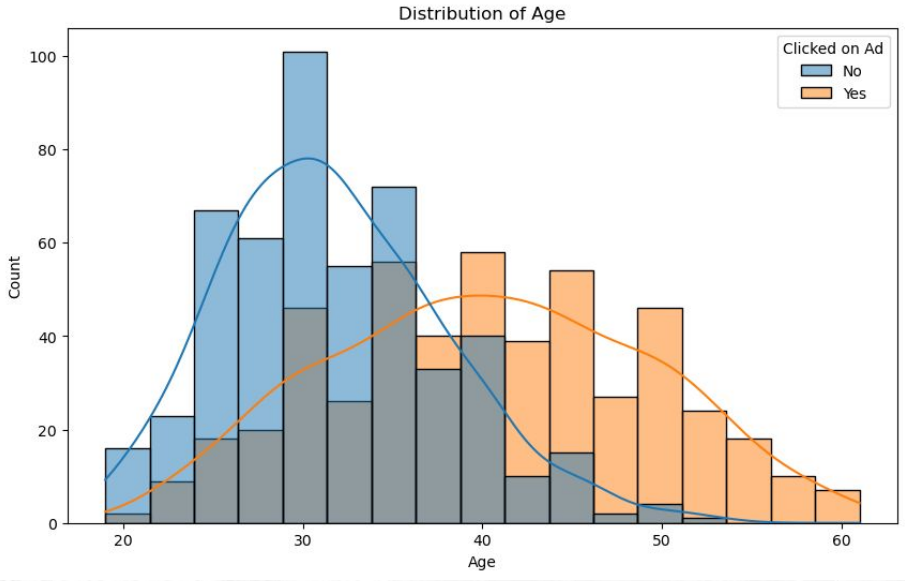
Statistical Analysis

- In terms of categorical features over advertisement, no significant difference observed.
  - Gender (p=.296)
  - Category (p=.695)
  - City (p=0.206)
  - Province (p=0.381)
- In terms of numerical features, age is having significant difference over advertisement (p=.000). However, other numerical features could not be tested due to missing data which will be addressed in the next stage.

For more information, see jupyter notebook here

EDA: Univariate Analysis (Age)

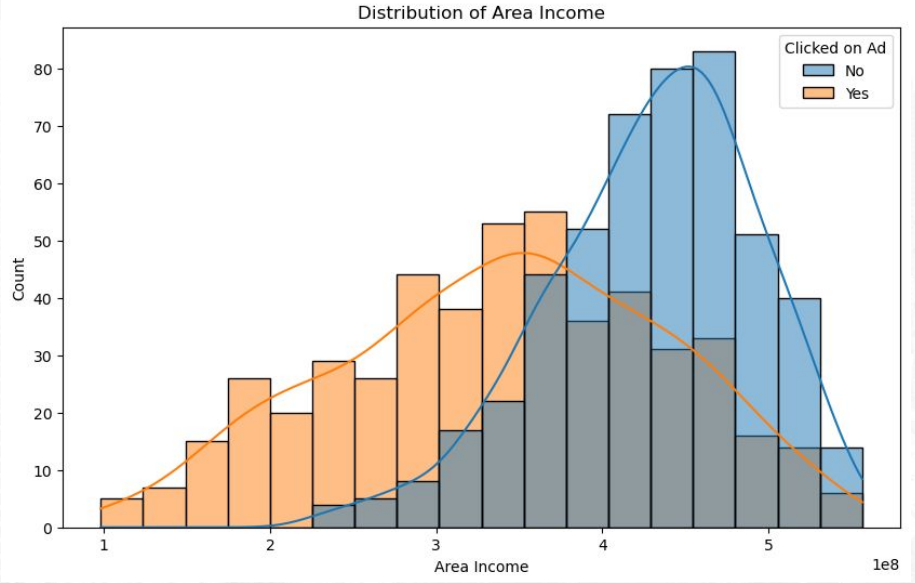
- Younger users (20-40) are less likely to click on ads, with a peak of non-clickers around 30 years old.
- Older users (40-60) are more likely to click on ads, with the highest proportion of clickers around 45-50 years old.
- This age distribution implies that ad strategies might be more effective when targeted towards older demographics.



For more information, see jupyter notebook here

EDA: Univariate Analysis (Area Income)

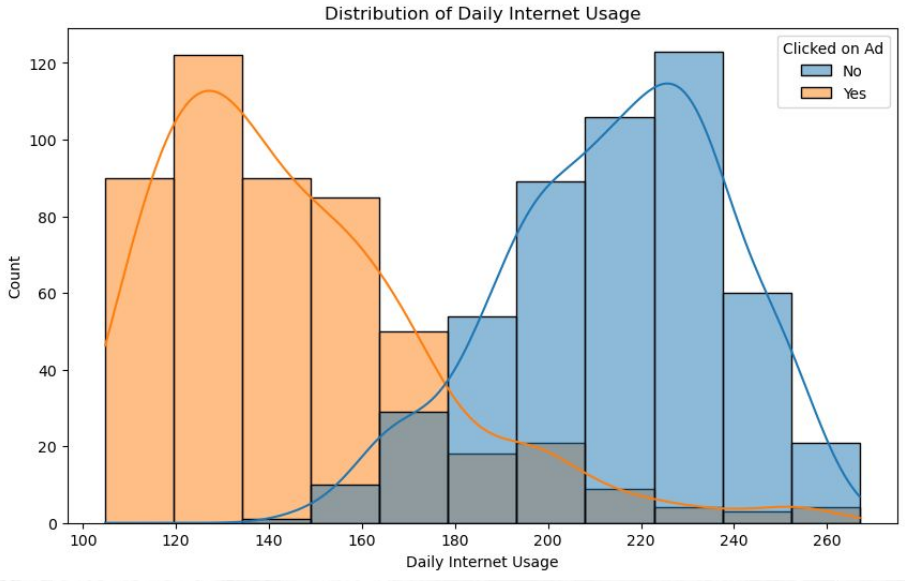
- Non-clickers (blue) are more prevalent in higher income areas, with a peak around 400-500 million.
- Ad-clickers (orange) are more common in lower to middle income areas, peaking around 200-300 million.
- This suggests that ads are more effective in lower to middle income areas, while higher income users are less likely to engage with ads.



For more information, see jupyter notebook here

EDA: Univariate Analysis (Daily Internet Usage)

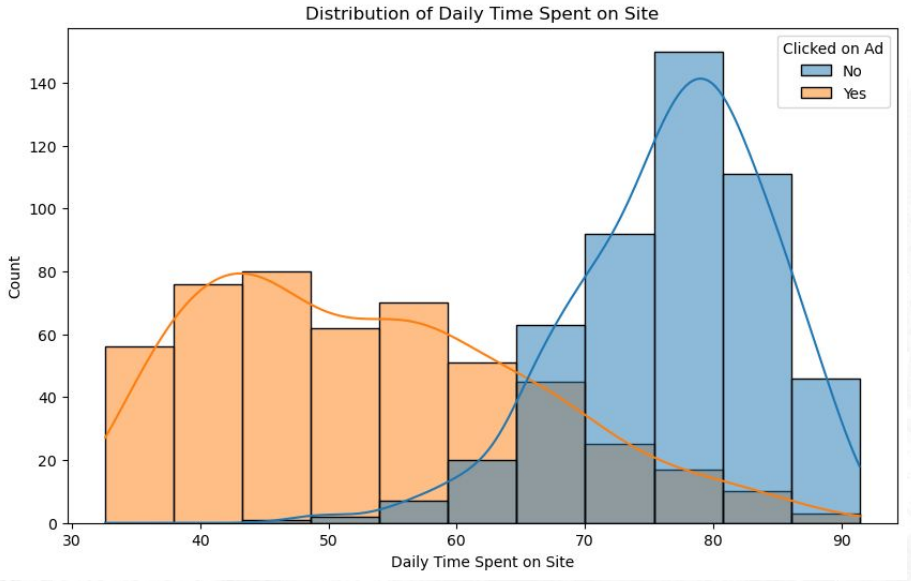
- Non-clickers (blue) have higher daily internet usage, mostly between 200-260 minutes.
- Ad-clickers (orange) show lower daily internet usage, concentrated between 100-160 minutes.
- This indicates that heavy internet users are less likely to click on ads, possibly due to ad fatigue or better ad recognition.



For more information, see jupyter notebook here

EDA: Univariate Analysis (Daily Time Spent on Site)

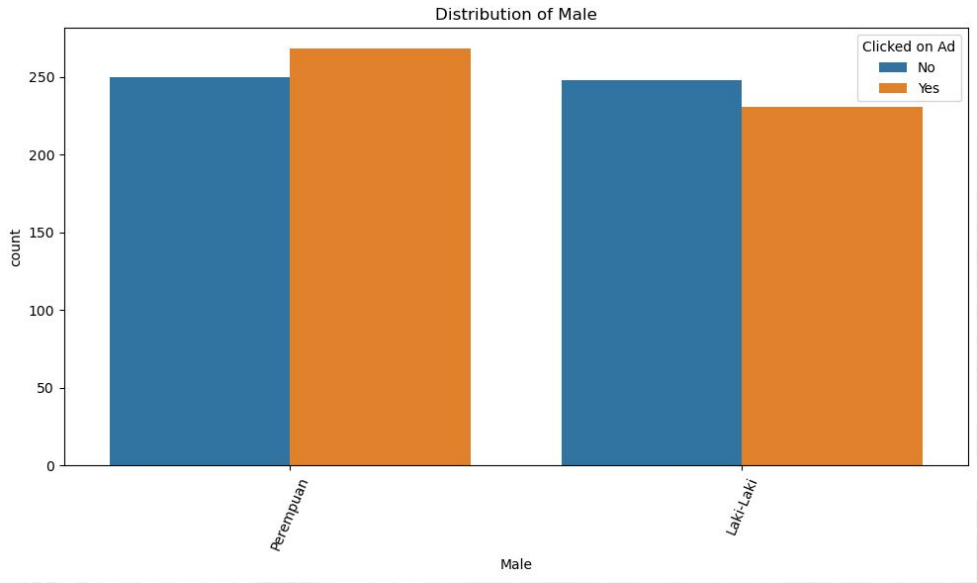
- Users who did not click on ads (blue) tend to spend more time on the site, with the majority spending between 70-90 minutes daily.
- Users who clicked on ads (orange) generally spend less time, with most spending between 30-60 minutes daily.
- There's a clear separation in behavior, suggesting that users who spend more time on the site are less likely to click on ads.



For more information, see jupyter notebook here

EDA: Univariate Analysis (Gender)

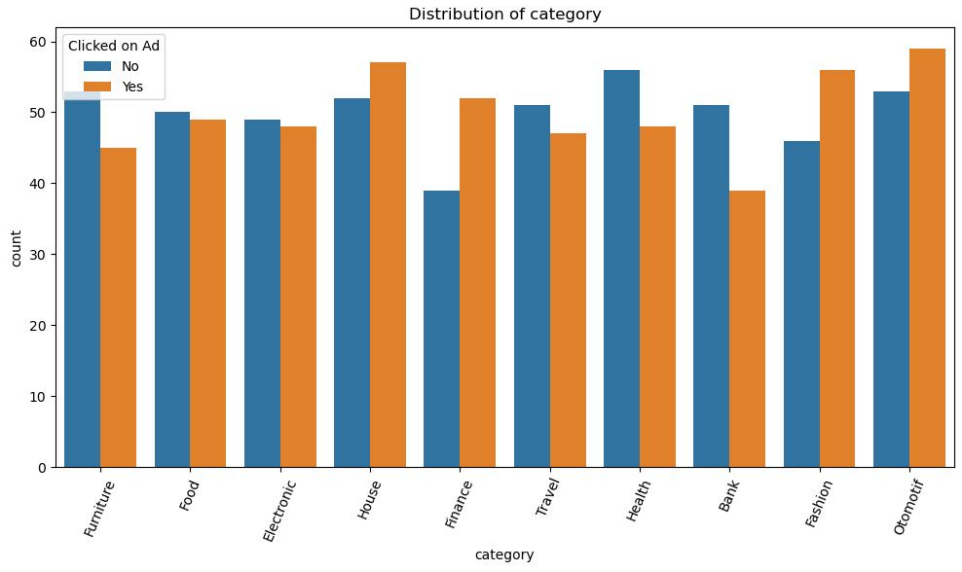
- There are two categories: "Perempuan" ("Female") and "Laki-laki" ("Male").
- "Perempuan" shows slightly higher overall numbers for both clicks and non-clicks.
- The difference between clicks and non-clicks is more pronounced in "Laki-laki".



For more information, see jupyter notebook here

EDA: Univariate Analysis (Category)

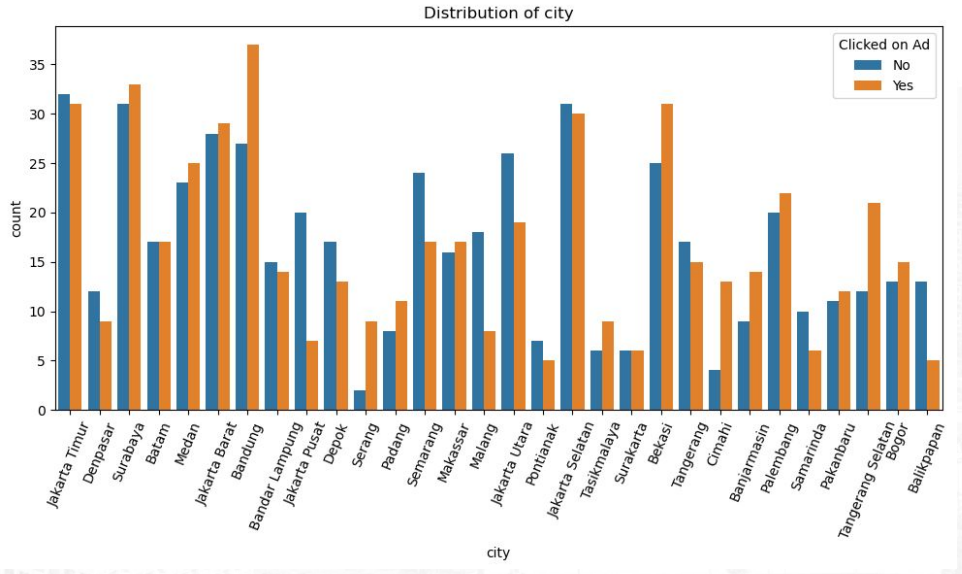
- The "Otomotif" category has the highest number of ad clicks, followed by "Fashion" and "House".
- "Furniture" and "Bank" categories have the lowest ad click rates.
- In most categories, the number of non-clicks (blue) is higher than clicks (orange), with notable exceptions in "House", "Fashion", “Otomotif” and "Finance".
- The "Health" and “Bank” category show the largest gap between clicks and non-clicks, with significantly more non-clicks.



For more information, see jupyter notebook here

EDA: Univariate Analysis (City)

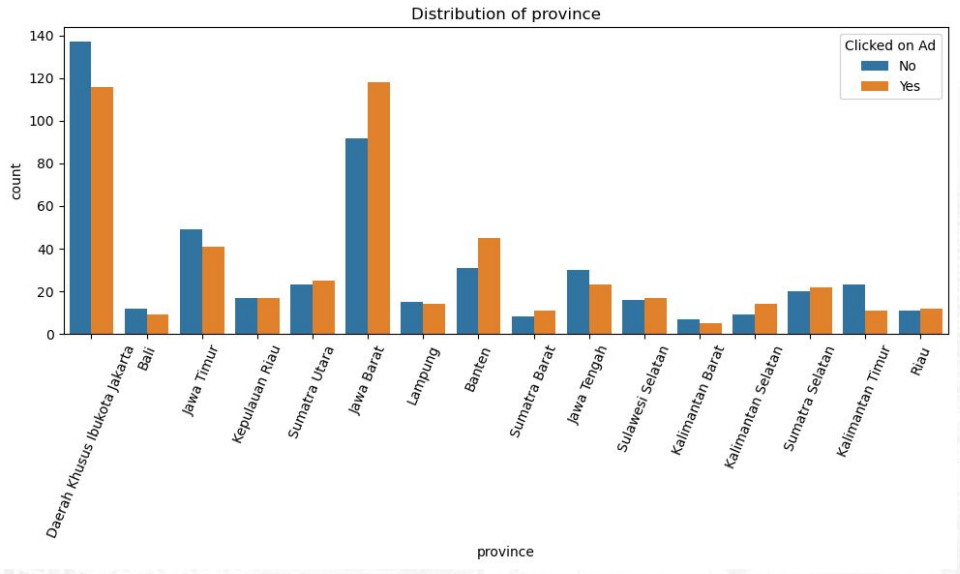
- Bandung has the highest number of clicks while Jakarta Timur with non-clicks.
- Cities like Jakarta Pusat, Jakarta Selatan, and Jakarta Timur also show high engagement.
- Some smaller cities or regions (e.g., Serang, Pekanbaru) have much lower overall counts.
- Click behavior varies significantly between cities, with some showing higher click rates relative to their population.



For more information, see jupyter notebook here

EDA: Univariate Analysis (City)

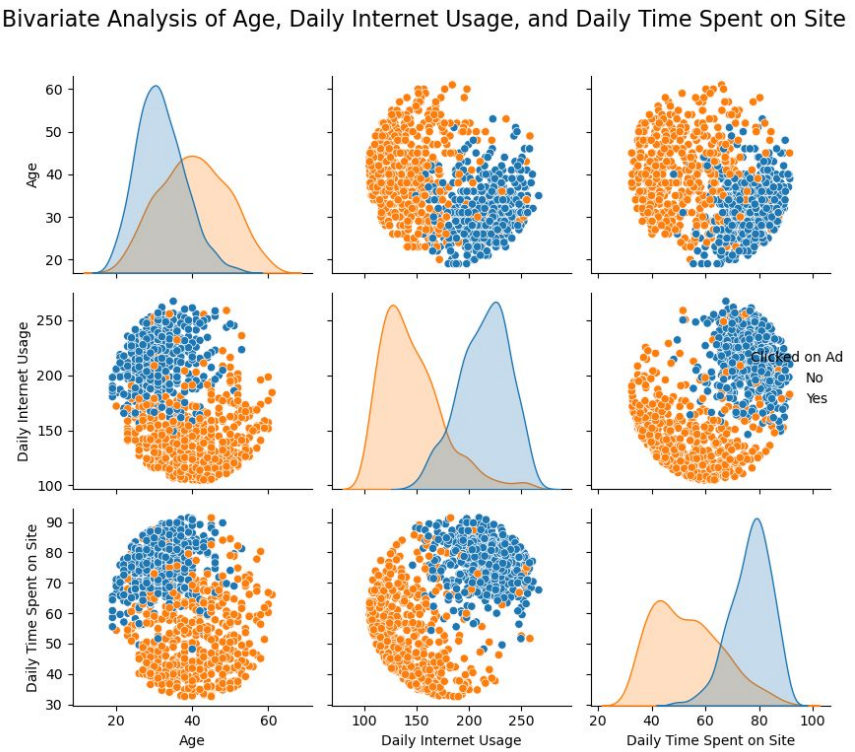
- DKI Jakarta has by far the highest number of both clicks and non-clicks.
- Jawa Barat and Banten also show significant engagement, but much less than DKI Jakarta.
- Many provinces have relatively low engagement numbers.
- The ratio of clicks to non-clicks varies between provinces, potentially indicating regional differences in ad effectiveness.



For more information, see jupyter notebook here

EDA: Bivariate Analysis (Age, Daily Internet Usage, Daily Time Spent on Site)

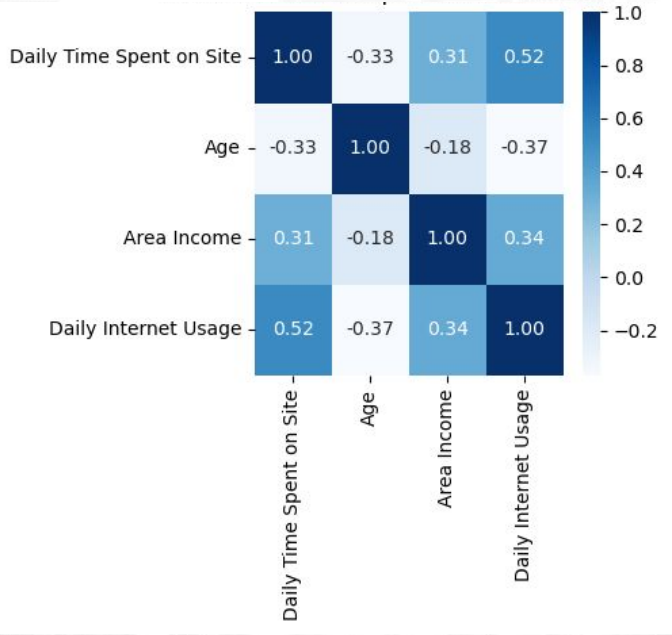
- Age vs. Daily Internet Usage: A weak positive correlation is observed between age and daily internet usage. As age increases, there's a slight tendency for daily internet usage to increase as well.
- Age vs. Daily Time Spent on Site: A very weak negative correlation is observed between age and daily time spent on site. Older individuals tend to spend slightly less time on the site compared to younger individuals.
- Daily Internet Usage vs. Daily Time Spent on Site: A moderate positive correlation is observed between daily internet usage and daily time spent on site. Individuals who use the internet more frequently tend to spend more time on the site.



For more information, see jupyter notebook here

EDA: Bivariate Analysis (Age, Daily Internet Usage, Daily Time Spent on Site)

- There is a strong positive correlation (0.52) between Daily Time Spent on Site and Daily Internet Usage, indicating that individuals who spend more time on the site tend to also use the internet more frequently.
- There is a weak negative correlation (-0.33) between Daily Time Spent on Site and Age, suggesting that older individuals tend to spend slightly less time on the site compared to younger individuals.
- There is a moderate negative correlation (-0.37) between Daily Internet Usage and Age, indicating that older individuals tend to use the internet slightly less frequently compared to younger individuals.
- Overall, these findings suggest that individuals who use the internet more frequently are also more likely to spend more time on the site, regardless of their age or area income. However, age and area income may have some minor influences on internet usage and time spent on site.



For more information, see jupyter notebook here