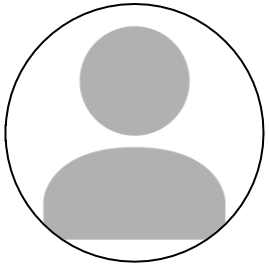
Predict Clicked Ads Customer Classification by Using Machine Learning

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

Rakamin Academy



Created by:
Bintang Phylosophie
bintangphy@gmail.com
https://www.linkedin.com/in/bintang-phylosophie/

A results-driven Data Scientist and Data Analyst with a strong foundation in machine learning, data analysis, and visualization. Proficient in Python, SQL, and advanced analytics tools such as Power BI, and Looker Studio. Experienced in handling large datasets, optimizing queries, and building data-driven solutions through project-based internships in several company. Holds a Data Science certification from Rakamin Academy with hands-on expertise in statistical modeling, predictive analytics, and interactive dashboard creation. Passionate about data storytelling and leveraging insights to drive business decisions.

Overview

Background

A company in Indonesia wants to know the effectiveness of the advertisement they broadcast. This is important for the company to know how successful the advertisement being marketed is 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.

Dataset & Business Understanding

Dataset Information:
This dataset appears to be about user behavior on a website, likely to explore how demographic and behavioral factors impact the likelihood of clicking on an ad. This dataset is from a fictional company from January to July in 2016.

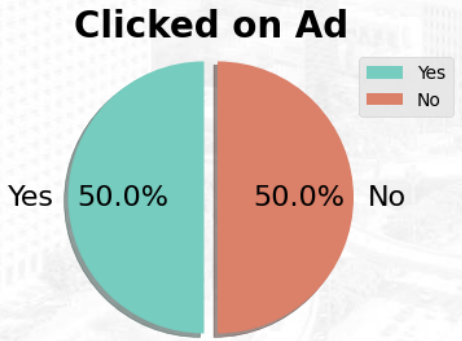
Attribute Information:

- **Identifier**
  This dataset does not contain an identifier, but it has a column called `Unnamed: 0`, which seems to function more as an index.
- **Target**
  In this case, the target variable, `Clicked on Ad`, is already available and will be used as the target, categorized as "**Yes**" and "**No**"
- **Company Goals:**
  - Optimizing ad spend by targeting relevant users.
  - Increasing ad engagement and conversion rates.
  - Segmenting customers based on their behavior and demographics.

.

- **Problem:**
  The company wants to evaluate the effectiveness of its advertisements in attracting customer engagement. Understanding which customers are more likely to click on ads will help improve targeted marketing efforts and optimize advertising spend. However, the company currently lacks a systematic approach to identifying potential customers who are most likely to interact with their ads.
- **Objectives:**
  - Identify key factors that influence ad clicks, such as time spent on the site, internet usage, age, income, and location.
  - Develop a predictive model to classify users into those who are likely to click on ads and those who are not.
  - Provide insights that can help the company refine its marketing strategies and improve ad targeting for better engagement.

What Happened?



Only half of user population clicked on the ad

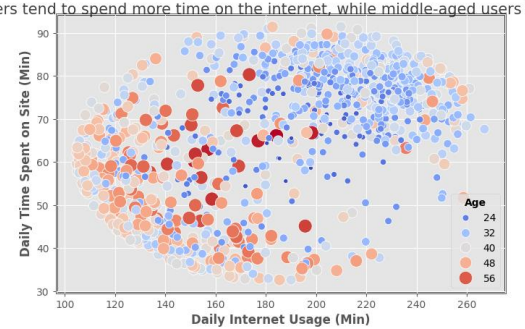
Click Here to see my code

Customer Type and Behaviour Analysis on Advertisement

Internet Usage

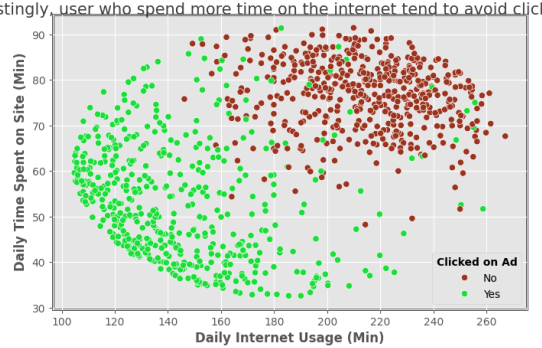
Relationship Between Daily Internet Usage vs Daily Time Spent on Site Based on Age

Younger users tend to spend more time on the internet, while middle-aged users spend less time.



Relationship Between Daily Internet Usage vs Daily Time Spent on Site Based on Clicking on Ad

Interestingly, user who spend more time on the internet tend to avoid clicking on ads



By analyzing users' age, daily internet usage, time spent on-site, and their willingness to click on ads, we can observe that:

- There is a **positive correlation** between internet usage and time spent on the site—users who spend more time online also tend to spend more time on site.
- The most concentrated region (high-density area) falls between **150 to 220 minutes of internet usage and 40 to 80 minutes on the site.**
- **Younger users** tend to spend more time online, while **middle-aged** users spend less time. In fact, middle-aged and older users are more likely to click on ads. Casual users (with lower internet usage) may be more curious and open to clicking on ads.
- However, spending more time on the internet **does not** necessarily make users more likely to click on ads. This suggests that users who are more active online tend to ignore ads, possibly due to ad fatigue or banner blindness.

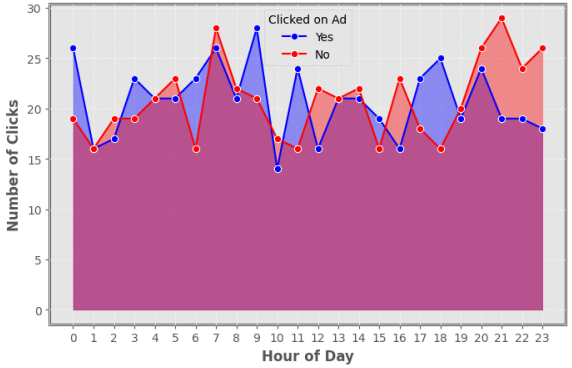
Click Here to see my code

Customer Type and Behaviour Analysis on Advertisement

Time Clicking on Ad

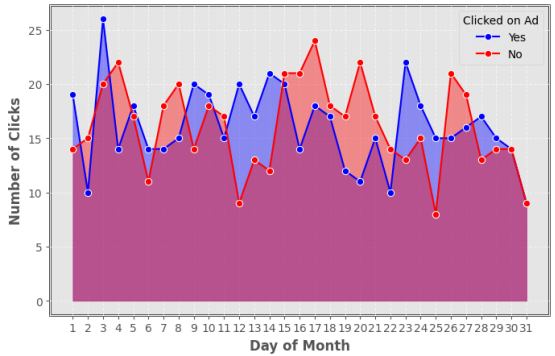
Number of Clicks on Ads by Hour of the Day

Ad clicks fluctuate throughout the day, with the highest number at 9 AM.
Users tend to avoid clicking ads between 9 and 11 PM.



Number of Clicks on Ads by Day of the Month

Ad clicks fluctuate throughout the month, with the highest clicks on the 3rd and the lowest on the 17th



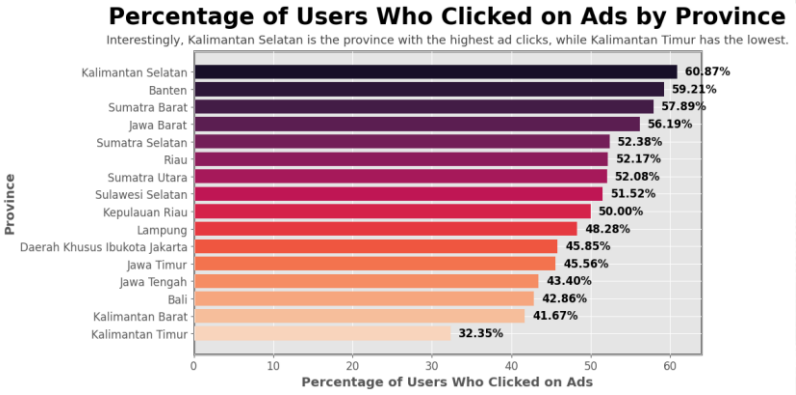
Looking at users' behavior on an hourly and daily basis, it fluctuates significantly, making it difficult to identify a clear pattern of when users are more likely to click on ads. However, some potential trends can still be observed:

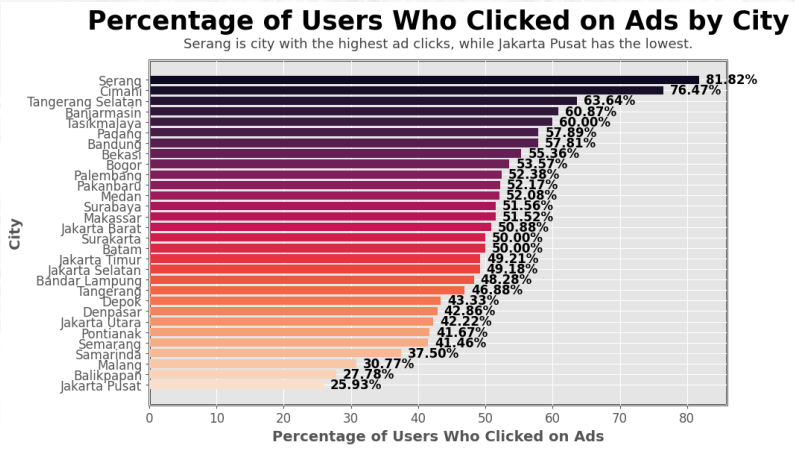
- Around 9 AM, 5 PM, and 9 PM, there are **higher "No" clicks**, suggesting users may be seeing ads but choosing not to interact with them. Users may be winding down for the day and ignoring ads.
- Between midnight and 6 AM, there are **slightly more "Yes" clicks**, indicating a higher engagement rate among night users. 9 AM Peak could be due to users checking emails, news, or social media early in the day. Running ads between 6 AM - 10 AM may yield the highest engagement.
- Ad clicks **do not follow a consistent pattern** and vary significantly across days.
- The highest number of clicks occurred on the **3rd day of the month**. Analyze the top-clicked days (e.g., 3rd and 23rd) to understand what marketing strategies were effective.
- The lowest number of clicks was recorded on the 17th day. More users ignored ads, such as the 17th and 18th days. **Investigate why the 17th had the lowest** engagement and adjust ad placements, timing, or targeting.

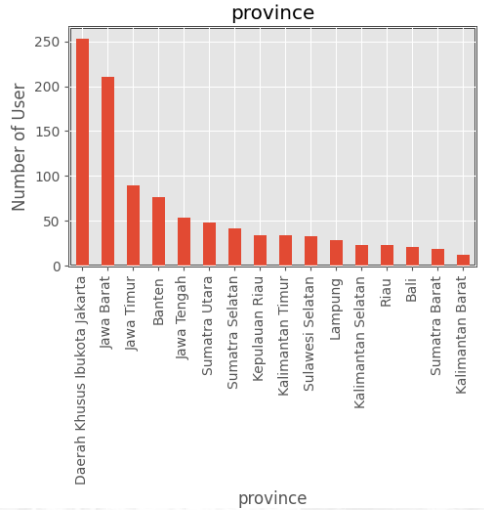
Click Here to see my code

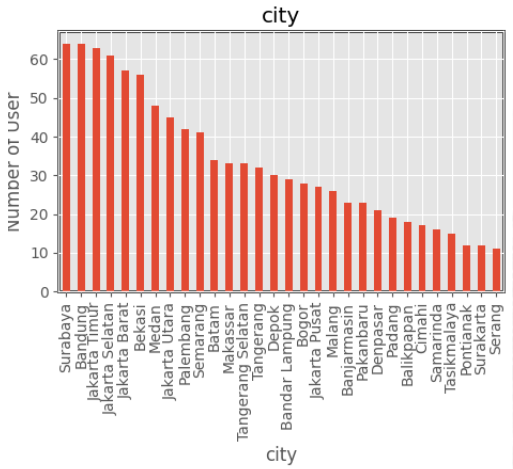
Customer Type and Behaviour Analysis on Advertisement

Geographic









Click Here to see my code

Customer Type and Behaviour Analysis on Advertisement

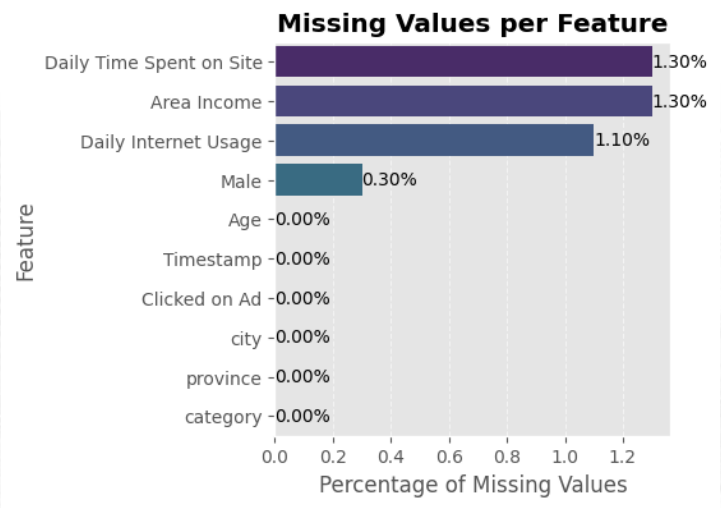
Rakamin Academy

Geographic

- The province-level and city-level **trends are consistent**, where cities with high engagement generally belong to provinces with high engagement.
- Kalimantan Selatan (60.87%) has **the highest percentage of ad clicks**, despite not having the highest user count, while Kalimantan Timur (32.35%) has **the lowest ad click rate**. This could be influenced by socioeconomic and behavioral differences. However, further investigation is needed, considering that the number of users in both cities does not differ significantly.
- Jakarta has a large audience but lower relative ad interaction. Jakarta Pusat, the lowest-ranking city (25.93%), aligns with other Jakarta-based areas also have moderate to lower rankings. **Urban areas like Jakarta Pusat tend to have lower ad-click rates**, possibly due to a different user behavior pattern or higher exposure to ads reducing their effectiveness.
- **Smaller cities like Serang (81.82%) and Cimahi (76.47%) have very high ad click rates despite lower user counts.** Possibly indicating higher ad relevance or a different level of online activity in those areas. Also users in smaller cities/provinces might have fewer digital distractions, leading to higher engagement.

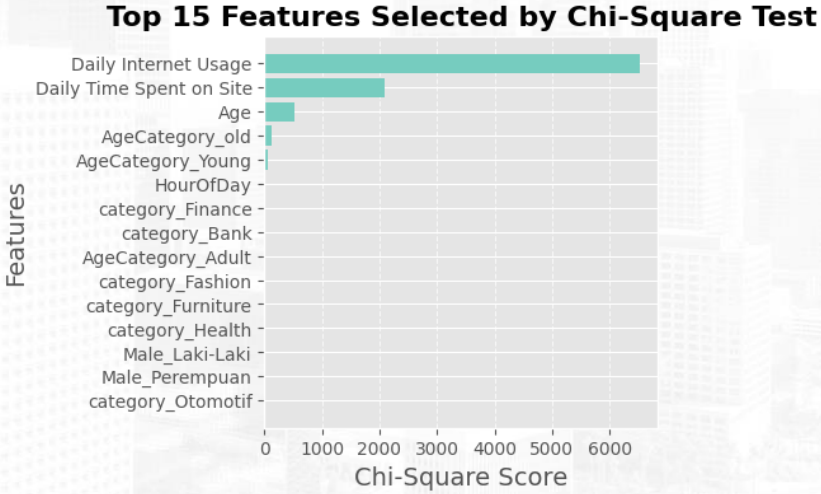
Data Preprocessing

Data Preprocessing



- When building machine learning models, we must address data skewness by **imputing missing values using the mean** for 'Daily Time Spent on Site' and 'Daily Internet Usage', and using **mode** for `Area Income` and 'Male'.
- Data distribution seems normal, we can conclude **that the dataset does not have outliers**..

Features Selection



- Since we are working with categorical features and a classification problem, **SelectKBest with chi2** is a great choice because it is: fast, helps remove irrelevant features, works well with encoded categorical data. Base on the result I will **choose top 5 features** for this machine learning model.
- **Label encoding** is applied to columns with higher cardinality, while **one-hot encoding** is used for columns with lower cardinality.
- Additionally, I perform scaling using **MinMaxScaler.**

Click Here to see my code

Modelling

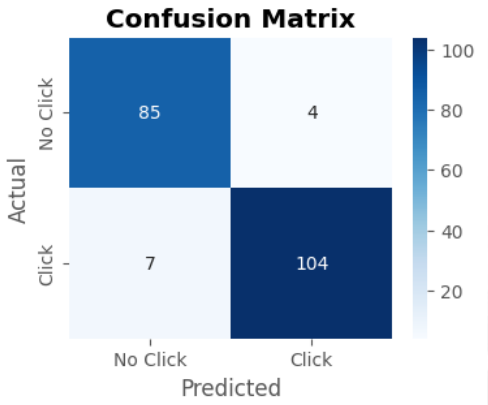
Model Development

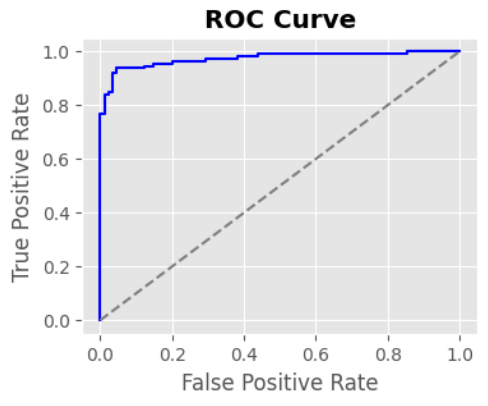
I use Support Vector Machine (SVM), Gradient Boosting, Decision Tree, Random Forest, and Logistic Regression.

Model Evaluation

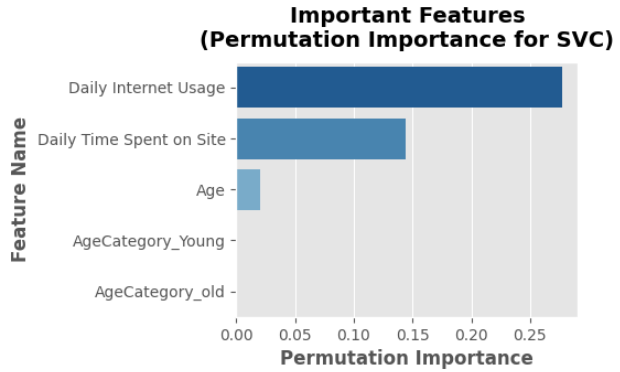
I choose **Precision** metrics Predict Clicked Ads Customer prediction to ensures that the model is optimized to minimize wasted ad spend on users who won't click.

Model	Precision	CV
Support Vector Machine	**0.96**	**0.98**
Gradient Boosting	0.94	0.96
Decision Tree	0.91	0.95
Random Forest	0.95	0.85
Logistic Regression	0.94	0.57





Feature Important



- Best model: **Support Vector Machine** with 10-folds cross validation
- The test was carried out using the Precision metric with Support Vector Machine model resulting Precision **0.96 and 0.98 after cross validation**
- Chart of Feature Importance showing that only the top 3 features significantly influence our model, indicating that we can focus on using those 3 features in understanding user behavior and optimizing ad campaigns.

Click Here to see my code

Business Recommendation

1. **Optimize Ad Spend & Targeting**
   Action: Focus advertising budgets on users with a high probability of conversion (e.g., those who spend more time on-site and have moderate internet usage).
   Implementation:
   - Use custom audience targeting in platforms like Google Ads & Facebook Ads.
   - Increase retargeting efforts for high-scoring users based on model predictions.

2. **Improve User Engagement for Low-Intent Visitors**
   Action: Action: Users who spend less time on-site but have high internet usage may need better engagement strategies to convert.
   Implementation:
   - Use exit-intent popups offering discounts for users with low engagement.
   - Optimize website content for better navigation and product discovery.

3. **Personalize Marketing for Different Age Groups**
   Action:
   Young Users (18-35) → Offer fast checkout, social proof, and influencer-driven campaigns.
   Older Users (50+) → Provide detailed product descriptions, trust-building elements, and customer support.
   Implementation:
   - A/B test different landing pages for different age segments.
   - Use dynamic pricing & promotions based on engagement patterns.

Business Simulation

A digital marketing agency, **AdTech Solutions,** partners with an e-commerce company to improve its online ad targeting strategy. The company aims to optimize its advertising budget by identifying users who are most likely to click on ads based on their demographic and behavioral characteristics.

Money Simulation: Optimizing Ad Revenue

Assumptions for Revenue Simulation:

- Cost per Click (CPC) = Rp 8,000 (each ad click generates Rp 8,000 revenue).
- Daily Ad Budget = Rp 160,000,000.
- Baseline Click-Through Rate (CTR) = 2% (without ML optimization).
- ML-Powered CTR Increase = 5% (after applying our predictive model).

Baseline Scenario (Without ML Optimization)

- Total Impressions per Day = 500,000
- CTR = 2%
- Total Clicks = 10,000
- Revenue = 10,000 × Rp 8,000
  = Rp 80,000,000
- Profit = Revenue - Ad Spend
  = Rp 80,000,000 - Rp 160,000,000
  = **- Rp 80,000,000 (Loss)**

Optimized Scenario (With ML Optimization)

- ML-Filtered Targeted Impressions = 300,000
- ML-Powered CTR = 5%
- Total Clicks = 15,000
- Revenue = 15,000 × Rp 8,000
  = Rp 120,000,000
- Profit = Revenue - Ad Spend
  = 120,000,000 - Rp 160,000,000
  = **- Rp 40,000,000 (lower loss)**

.

Break-Even Point Calculation

To break even:

- Required Clicks = Rp 160,000,000 /Rp 8,000 = 20,000 clicks
- Required CTR = 20,000 / 300,000 = 6.67%
- By further improving the model, segmenting users better, and using A/B testing, the company can push CTR beyond 6.67% and start making a profit

Business Insights & Takeaways

- ML-based ad targeting reduces wasted impressions and increases CTR from 2% to 5%.
- More clicks from relevant users lead to higher revenue (Rp 120,000,000
- instead of Rp 80,000,000).
- Better audience segmentation (e.g., high-income users for luxury ads) maximizes engagement.
- Further improvements (A/B testing, reinforcement learning) can push CTR above 6.67% to reach profitability.
- Long-term gain: As models improve, more precise targeting will lead to higher ROI and reduced costs.