Predict Clicked Ads Customer Classification by using Machine Learning

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

Created by:
Bagus Ariobimo
bagusariobimo7@gmail.com
https://www.linkedin.com/in/bagus-ariobimo
https://github.com/riyouuyt

“A data enthusiast who has completed a course in this field and is ready to start his career. Have a excellent understanding in statistics, programming, and data processing. Proficient in using tools such as Python, R, and SQL. Able to collect, clean and analyze data with the necessary techniques. Have skills in data visualization and simple statistical modeling. Creative problem solver and always thrive for learning. Ready to contribute to data-driven projects and collaborate in teams. Committed to further developing data science skills and achieving significant results in data analysis.”

Overview

Project Overview

A company in Indonesia wants to evaluate the effectiveness of the advertisements they broadcast. This is crucial for the company to assess the success of their marketing efforts in attracting customers to view the advertisements. By analyzing historical advertising data and identifying insights and patterns, this project aims to assist the company in defining their marketing targets. The primary objective of this project is to create a machine learning classification model that can determine the right customer targets.

Goals

- Create Effective Model: Build a machine learning model to predict ad clicks accurately.
- Enhance Marketing Strategy: Use data insights to improve the company's marketing campaigns.
- Optimize Targeting: Help the company target the right audience for better ad performance.

Objectives

- Model Development: Develop a robust ad click prediction model.
- Insight Generation: Identify actionable insights from historical ad data.
- Data-Driven Decisions: Promote data-driven marketing decisions within the company.

Data Preparation

📚 Python Libraries for The Project:

pandas 🐼: for data manipulation.
numpy 🔢: for numerical operations.
seaborn 🌈: for data visualization.
matplotlib 📊: for plotting.
scikit-learn ✖ **:** for machine learning tasks.

🤖 Machine Learning Goals:

- The goal is to predict whether a user will click on an ad based on their behavior and demographics.
- Feature engineering and model selection will be used to achieve this goal.

📊 Dataset Overview:

Total Rows: 1000
Total Columns: 11

Columns Explanation:

- **’Daily Time Spent on Site’**: Average time spent by users on the website.
- **’Age’**: Age of users.
- **’Area Income’**: Income level of users' areas.
- **’Daily Internet Usage’**: Daily internet usage by users.
- **’Male’**: Gender of users (binary: "Male" or "Female").
- **’Timestamp’**: Timestamps or dates of user data.
- **’Clicked on Ad’**: Indicator of whether a user clicked on an ad (binary: "Yes" or "No").
- **’City’**: City where users are located.
- **’Province’**: Province or region of users.
- **’Category’**: Categorical variable or category.

Exploratory Data Analysis

NUMERICAL FEATURES 🔢

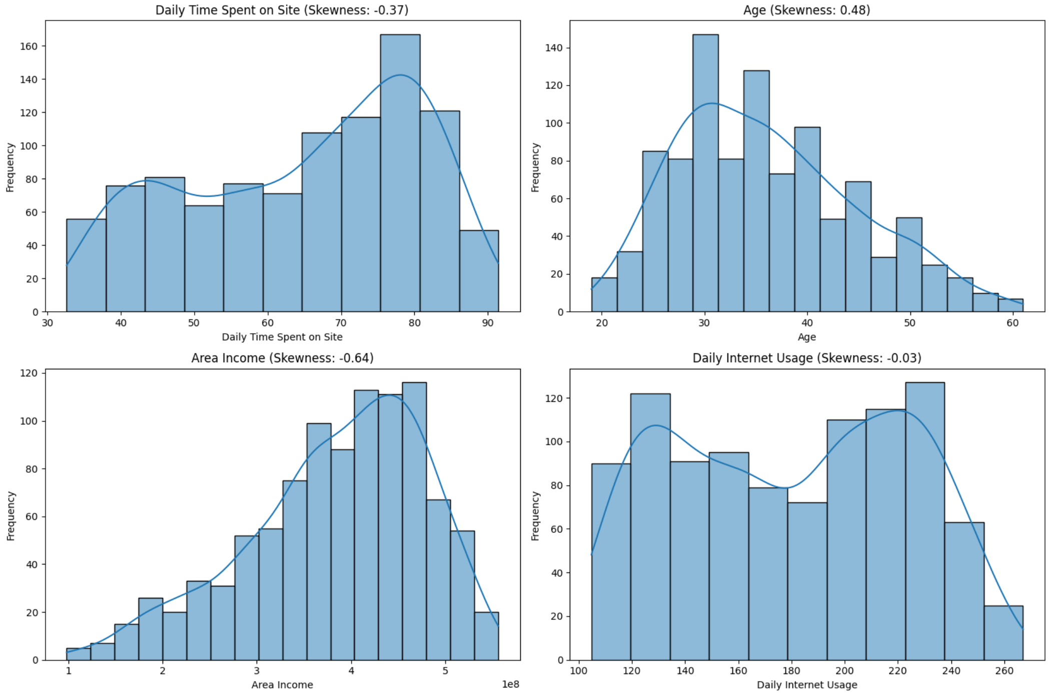
☑ KEY INSIGHTS BASED ON HISTOGRAM:

1. **DAILY TIME SPENT ON SITE:**
   - NEARLY SYMMETRICAL DISTRIBUTION.
   - SLIGHT NEGATIVE KURTOSIS INDICATES A FLATTER DISTRIBUTION.

2. **AGE:**
   - SLIGHTLY POSITIVELY SKEWED DISTRIBUTION.
   - NEGATIVE KURTOSIS SUGGESTS A RELATIVELY FLAT PEAK.

3. **AREA INCOME:**
   - MODERATELY NEGATIVELY SKEWED DISTRIBUTION.
   - KURTOSIS CLOSE TO ZERO INDICATES A NORMAL-LIKE DISTRIBUTION.

4. **DAILY INTERNET USAGE:**
   - NEARLY SYMMETRICAL DISTRIBUTION.
   - NEGATIVE KURTOSIS INDICATES A FLAT PEAK.



For more info of the project, click here

Exploratory Data Analysis

NUMERICAL FEATURES

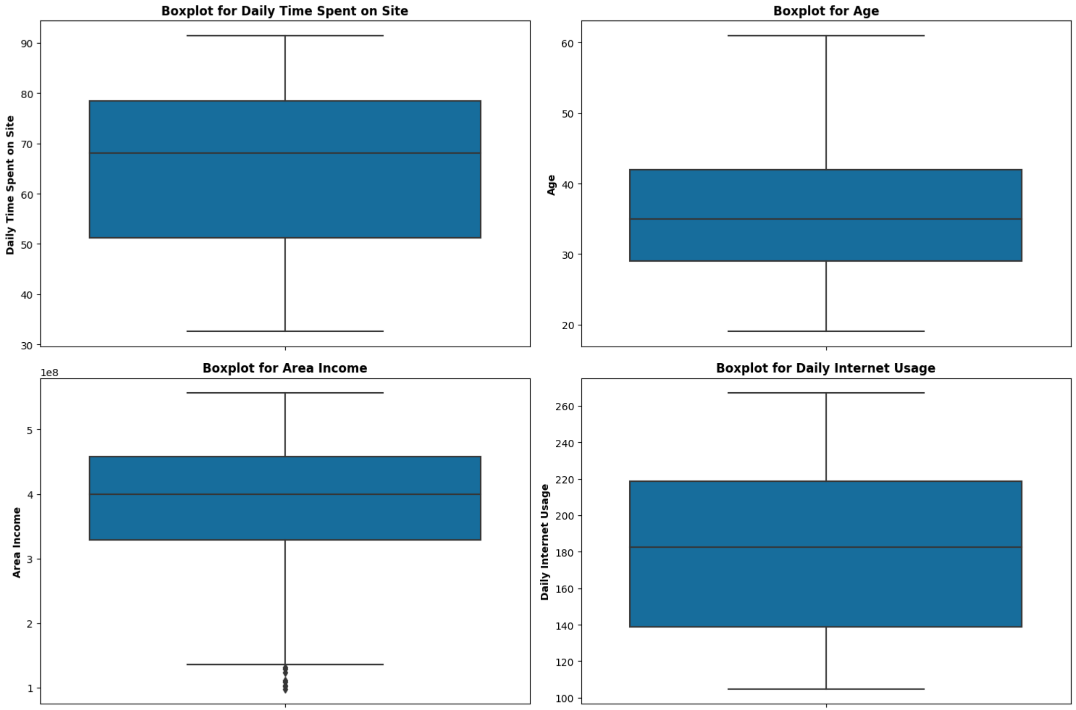
Key Takeaways:

- For the feature Daily Time Spent on Site, the data exhibits a spread with an Interquartile Range (IQR) of 27.19. Users tend to spend varying amounts of time on the site, with the middle 50% spending between 51.27 and 78.46 units.

- In terms of Age, there's a notable spread in user ages, as indicated by an IQR of 13.00. The middle 50% of users fall between the ages of 29 and 42

- The feature Area Income displays significant variability, with an IQR of 129,722,500.00. Users' incomes in the middle 50% range from 328,633,000.00 to 458,355,400.00, indicating diverse income levels.

- Daily Internet Usage demonstrates a wide range of usage, with an IQR of 80.08. Users in the middle 50% consume daily internet services between 138.71 and 218.79 units.



For more info of the project, click here

Exploratory Data Analysis

CATEGORICAL FEATURE

🏙 Province Distribution:

The top 5 provinces with the highest representation among users are:

- **Daerah Khusus Ibukota Jakarta** (253 users)
- **Jawa Barat** (210 users)
- **Jawa Timur** (90 users)
- **Banten** (76 users)
- **Jawa Tengah** (53 users)

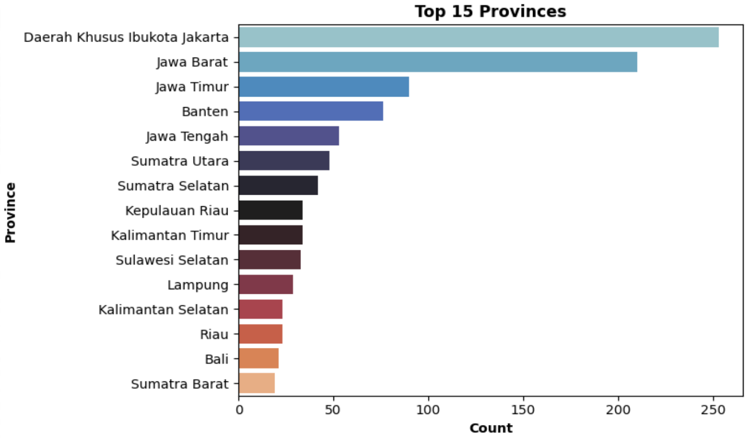
Jakarta's capital region, Daerah Khusus Ibukota Jakarta, has the highest user count.

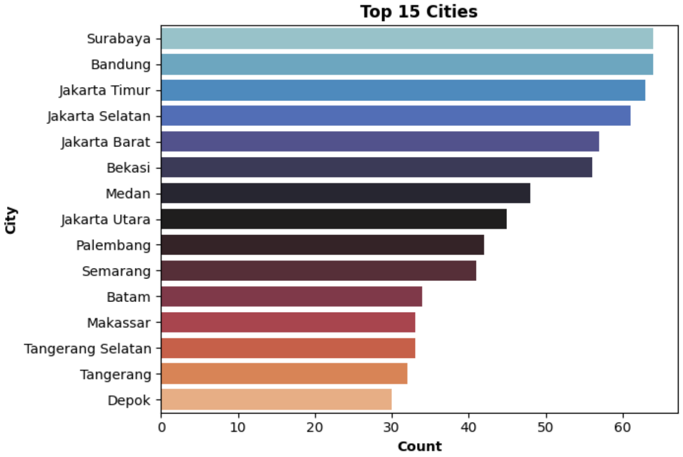
🏢 City Distribution:

The top 5 cities with the highest representation among users are:

- **Surabaya** (64 users)
- **Bandung** (64 users)
- **Jakarta Timur** (63 users)
- **Jakarta Selatan** (61 users)
- **Jakarta Barat** (57 users)

Surabaya and Bandung are the most prominent cities among users.





Exploratory Data Analysis

Rakamin Academy

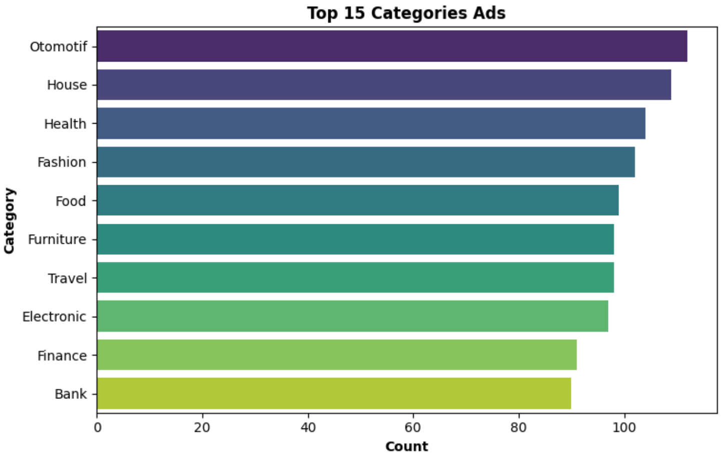
CATEGORICAL FEATURE

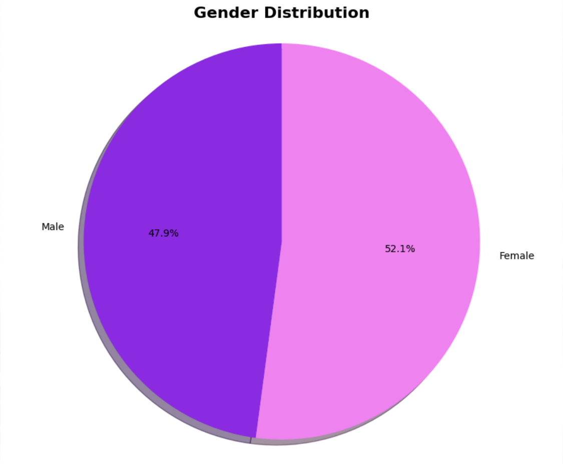
Ads Categories 📢

We examined the distribution of users among different ad categories. The top ad categories include **'Otomotif,' 'House,' 'Health,' 'Fashion,' 'Food,' 'Furniture,' 'Travel,' 'Electronic,' 'Finance,'** and **'Bank.'** Understanding these categories helps us tailor our marketing strategies to target specific interests and preferences of our users.

Gender Distribution 👫

We analyzed the gender distribution among our users. Approximately 47.9% of users identify as **Male**, while 52.1% identify as **Female**. Additionally, there were 3 cases (0.3%) with missing gender information. This gender breakdown allows us to customize our marketing campaigns to resonate with different segments of our audience.

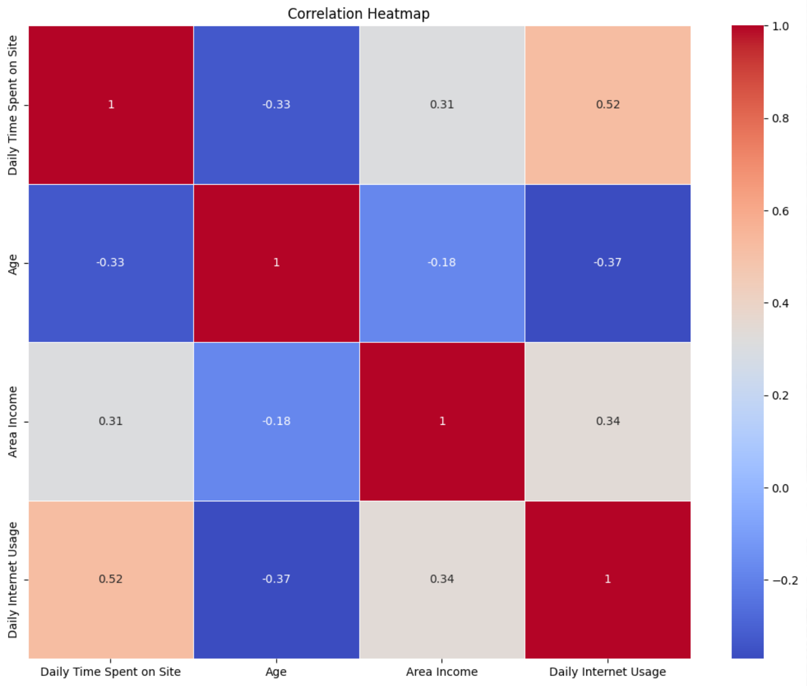


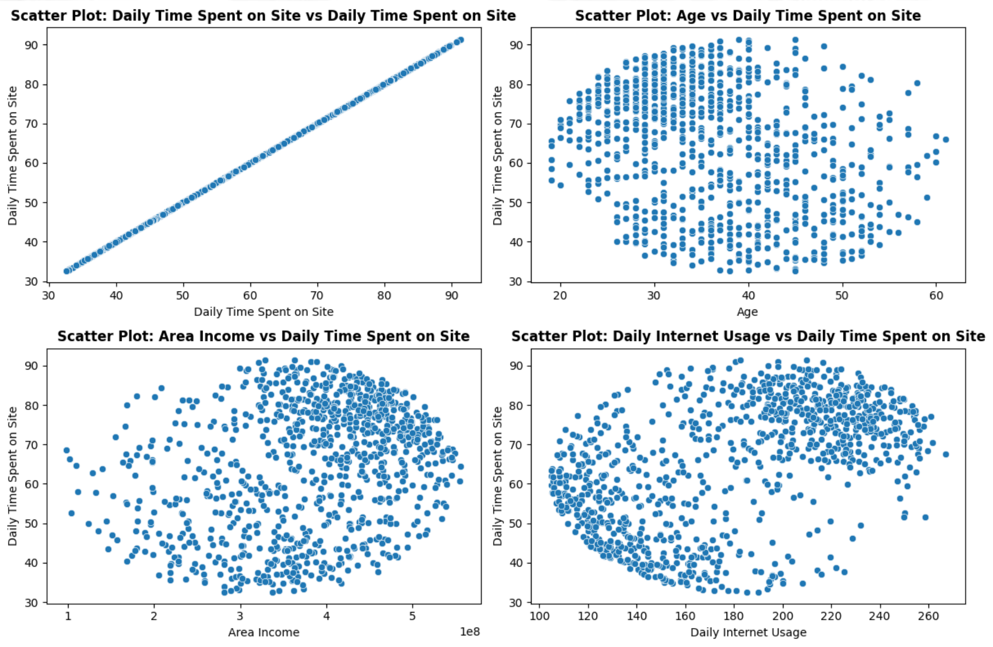


Exploratory Data Analysis

Rakamin Academy

BIAVARIATE ANALYSIS





For more info of the project, click here

Exploratory Data Analysis

BIAVARIATE ANALYSIS

Summary for the all the correlation columns above:

- **Daily Time Spent on Site vs. Age:** Negative correlation (-0.3314). Younger users tend to spend more time on the site.
- **Daily Time Spent on Site vs. Area Income:** Positive correlation (0.3083). Users from higher-income areas spend slightly more time on the site.
- **Daily Time Spent on Site vs. Daily Internet Usage**: Strong positive correlation (0.5183). More time on the site is associated with higher daily internet usage.
- **Age vs. Area Income:** Negative correlation (-0.1793). Younger users often reside in lower-income areas.
- **Age vs. Daily Internet Usage:** Negative correlation (-0.3705). Younger users have higher daily internet usage.
- **Area Income vs. Daily Internet Usage**: Positive correlation (0.3381). Higher area income relates to higher daily internet usage.
- **Strongest Correlation:** Daily Time Spent on Site and Daily Internet Usage (0.5183). A strong positive relationship exists.
- **Age vs. Daily Time Spent on Site:** Negative correlation (-0.3314). Older users spend less time on the site.

In summary, user engagement is positively linked to internet usage, with younger users being more active. Additionally, users from higher-income areas tend to engage more with the site. However, age and area income are negatively related, suggesting that younger users often reside in lower-income areas. These insights can inform marketing strategies and target demographic

Customer Type and Behaviour Analysis on Advertisement

📈 Customer Type and Behavior Analysis on Advertisement 📈

In our comprehensive analysis, we delved into customer behavior patterns and demographics, shedding light on their interaction with digital advertisements.

- Ads Categories 📢: We examined the distribution of users across various ad categories, helping us understand their preferences and interests. The top categories include **'Otomotif,' 'House,' 'Health,' 'Fashion,' 'Food,' 'Furniture,' 'Travel,' 'Electronic,' 'Finance,' and 'Bank.'** This knowledge empowers us to tailor our marketing strategies to align with specific customer interests.

- Gender Distribution 👫: Our analysis revealed a balanced gender distribution, with approximately 47.9% of users identifying as Male and 52.1% as Female. There were 3 cases (0.3%) with missing gender information. This insight enables us to create targeted and inclusive marketing campaigns.

- Customer Type 🎯: Through behavioral analysis, we identified different customer types based on their interaction with advertisements. By categorizing users into various segments, we gain a deeper understanding of their preferences, needs, and responsiveness to our marketing efforts.

- Age Groups 🎂: We explored age intervals to determine which age ranges are more likely to engage with advertisements. This information aids in crafting age-specific marketing strategies for better results.

For more info of the project, click here

Data Cleaning & Preprocessing

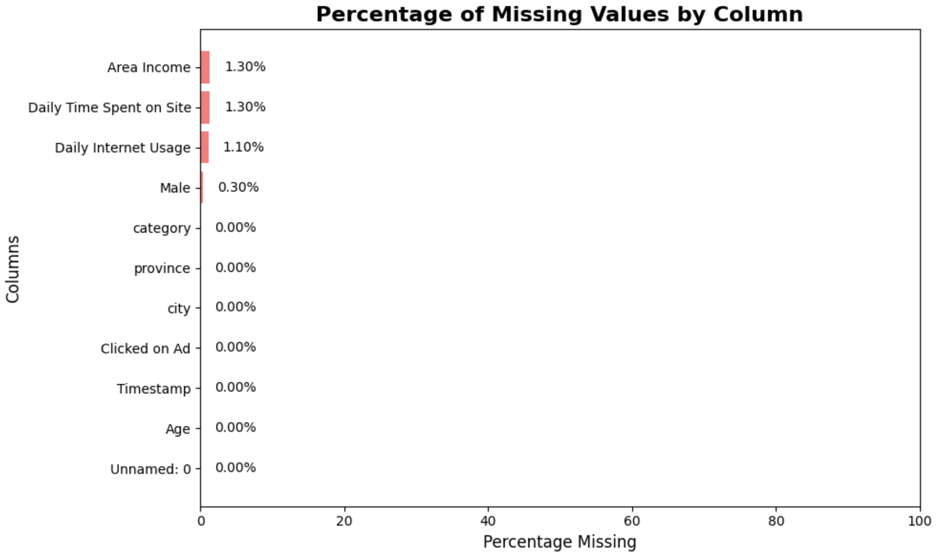
Checking Null Values in Dataset 🕵

In our dataset, we meticulously examined the presence of missing values. Here's a quick overview:

📊 Columns with Null Values:

- **'Daily Time Spent on Site' and 'Area Income'** each had a 1.3% null rate.
- **'Daily Internet Usage'** had a 1.1% null rate.
- **'Male'** exhibited a mere 0.3% of missing values.
- **'Unnamed: 0,' 'Age,' 'Timestamp,' 'Clicked on Ad,' 'city,' 'province,'** and **'category'** were entirely free from null values.

Ensuring data integrity is essential for reliable analysis. We'll handle these missing values appropriately to maintain data quality. 🛠



For more information about the project, click here

Data Cleaning & Preprocessing

□ Data Cleaning and Preprocessing Summary □

Our data underwent a comprehensive cleaning and preprocessing journey, ensuring its readiness for analysis. Here's what we accomplished:

- 🚮 **Handling Missing Values:** We meticulously addressed missing values, ensuring minimal data loss: **'Daily Time Spent on Site,' 'Area Income,' 'Daily Internet Usage,'** and **'Male'** columns had missing values, which were handled appropriately. **'Unnamed: 0,' 'Age,' 'Timestamp,' 'Clicked on Ad,' 'city,' 'province,' and 'category'** columns remained free from null values.
- □ **Duplicate Data:** We identified and successfully removed duplicate rows from the dataset, leaving us with a pristine dataset, completely devoid of duplicates.
- 🔄 **Feature Encoding:** We encoded categorical columns **'Gender'** and **'Clicked on Ad'** using one-hot encoding, making them ready for machine learning models.
- 📛 **Column Renaming:** We renamed columns for clarity and consistency, including renaming **'Gender'** columns for better interpretation.
- 🔪 **Column Removal:** We removed the **'Unnamed: 0'** column, which didn't contribute to our analysis.

Now, with our dataset clean, structured, and encoded, we're primed for in-depth analysis and predictive modeling! 🚀🔍

For more information about the project, click here

Data Splitting and Date Extraction Summary

In our data preprocessing journey, we took crucial steps to prepare the dataset for analysis:

- 🎯 **Splitting Data:** We divided our dataset into features (X) and the target variable (y). The **'Clicked on Ad_Yes'** column serves as our target, while the remaining columns comprise our features.

- ⏰ **Extracting Date Information:** To unlock valuable insights, we extracted date-related details from the **'Timestamp'** column:
  - We converted the **'Timestamp'** column to datetime format to ensure accurate time-related calculations.
  - We created new columns, including **'Year,' 'Month,' 'Week,' and 'Day,'** to dissect the timestamp data into various time components.

These meticulous preparations set the stage for our exploratory data analysis and predictive modeling, empowering us to uncover meaningful patterns and make informed decisions! 🚀 📈

For more information about the project, click here

Data Modeling

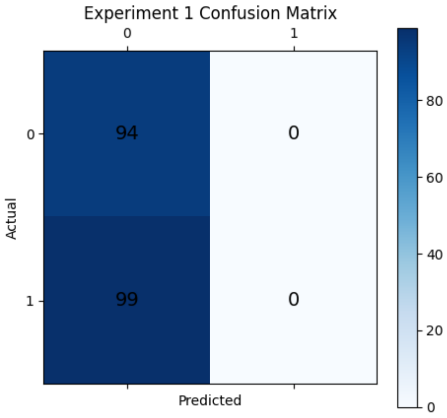
Machine Learning Experiments:

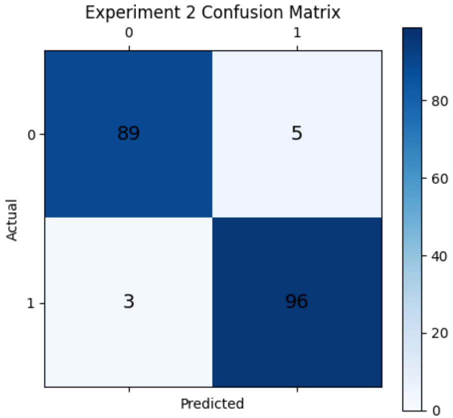
1. **Experiment 1 (Before Normalization/Standardization):**

- Low accuracy (49%) and poor performance.
- Top Features: Daily Time Spent on Site, Daily Internet Usage.

2. **Experiment 2 (After Normalization/Standardization):**

- High accuracy (96%) and improved performance.
- Top Features: Daily Time Spent on Site, Daily Internet Usage.



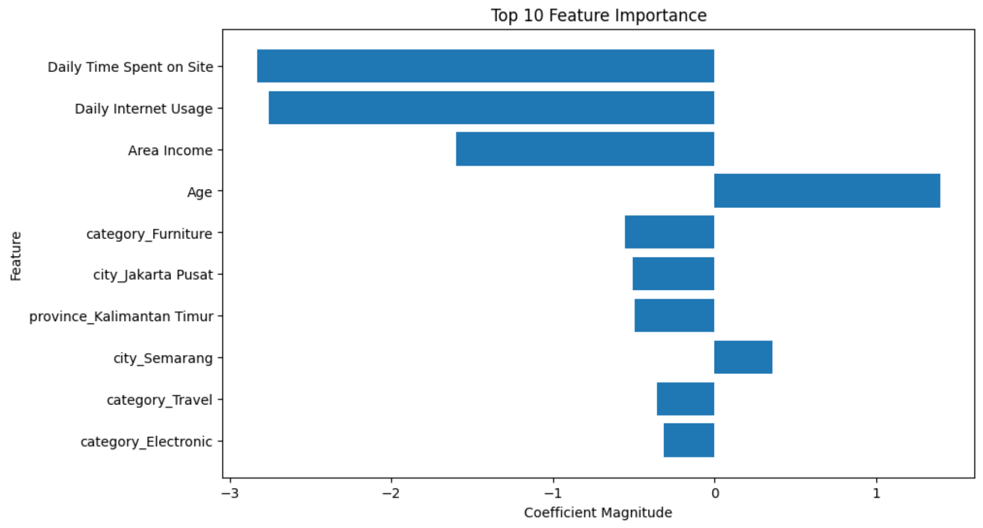


For more information about the project, Click Here

Data Modeling

Top Features by Coefficient Magnitude:

- Daily Time Spent on Site: -2.83
- Daily Internet Usage: -2.76
- Area Income: -1.60
- Age: 1.40
- category_Furniture: -0.56
- city_Jakarta Pusat: -0.51
- province_Kalimantan Timur: -0.50
- city_Semarang: 0.36
- category_Travel: -0.35
- category_Electronic: -0.32



These features have the highest absolute coefficient magnitudes in your logistic regression model. Features with negative coefficients have a negative impact on the prediction of clicking on an ad, while features with positive coefficients have a positive impact.

For more information about the project, Click Here

Summary

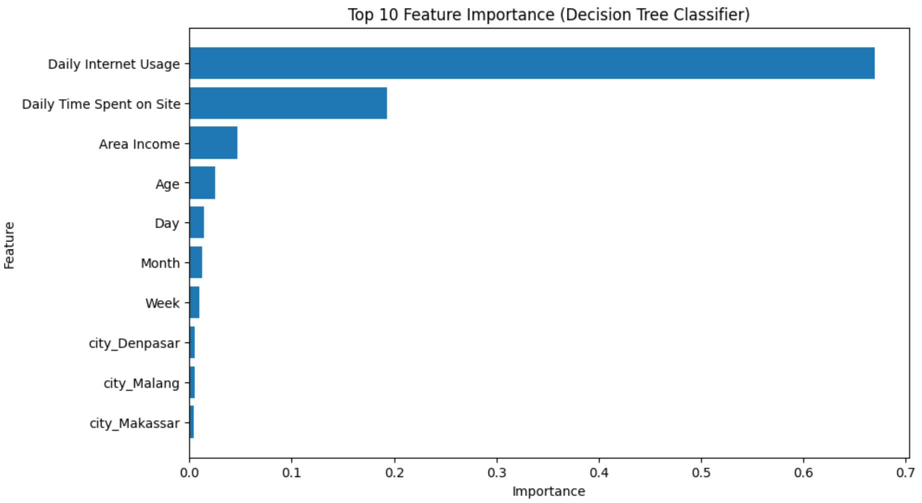
📈 Interpretation of Model Results □

- Experiment 1 lacked data normalization and performed poorly in predicting ad clicks. The top features, Daily Time Spent on Site and Daily Internet Usage, were not sufficient for accurate predictions.
- Experiment 2, with data normalization, significantly improved accuracy and precision. It successfully identified potential ad clickers.
- The top features affecting predictions are **Daily Time Spent on Site** and **Daily Internet Usage**, highlighting their importance in understanding user behavior.
- Data normalization plays a crucial role in enhancing model performance, making it a recommended practice.

Use these insights to refine marketing strategies and target potential ad clickers effectively. 🚀

For more information about the project, Click Here

Business Recommendation & Simulation



The feature importance analysis has identified that **"Daily Internet Usage"** and **"Daily Time on Site"** are the two most influential factors affecting user behavior and ad click-through rates. 🚀

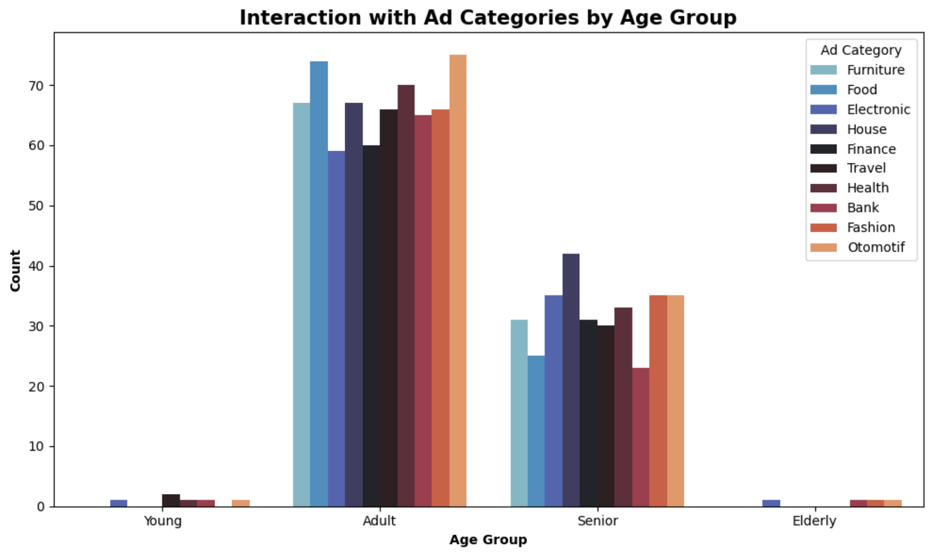
- **Daily Internet Usage**: This feature is a critical determinant of user engagement. Users who spend more time online tend to be more active and attentive. This is a strong indicator of their responsiveness to ads.

- **Daily Time on Site**: The time users spend on the site directly impacts their exposure to ads. Longer visits provide more opportunities for users to interact with advertisements.

These two factors, **"Daily Internet Usage"** and **"Daily Time on Site,"** are highly valuable in understanding user behavior and optimizing ad campaigns. 📊💡

For more information about the project, Click Here

Business Recommendation & Simulation



Age Group Distribution📊

Based on the feature importance, The age is pretty importance and can add some recommendation into the business distribution is predominantly concentrated in the "Adult" and "Senior" categories, with these two groups making up nearly the entire user base. The "Young" and "Elderly" groups represent only a small portion of the users.

Business Recommendation 🚀

- Target Adult and Senior Groups: 🎯 The marketing strategy should be primarily designed to target the "Adult" and "Senior" age groups since they make up the majority of users. Tailor advertisements, promotions, and content to resonate with the interests and preferences of these age groups.

- Explore Opportunities: 🔍 While the "Young" and "Elderly" groups are smaller in number, they still represent potential market segments. It may be worth exploring specific strategies to attract and engage these demographics, considering their unique interests and needs.

This business recommendation aims to optimize marketing efforts by focusing on the most significant user segments while exploring potential growth opportunities. 📈

For more information about the project, Click Here

BUSINESS SIMULATION

💰*Business Schemas*💰

Company Profile: CarMend Pro 🚗🔧

CarMend Pro is a customer-focused automotive services company specializing in car maintenance and repair. We connect car owners with trusted mechanics, offering convenient and reliable solutions to keep their vehicles in top condition. Our transparent pricing, user-friendly platform, and dedication to innovation make us a leader in the automotive industry.

Marketing Team Target: 🎯

- Objective:

To drive traffic to the CarMend Pro website, generate interest in our services, and increase sales. 🚗💻📈

- Target Audience: 🎯

Car owners looking for automotive maintenance and repair services. 🚗🔧💡

For more information about the project, Click Here

Assumed Clicked Customers🖱

Before the machine learning for the classification customers exist, the marketing teams are assuming the prospect customer which there are:

- **Interested Prospects (Assumed 30%):** □
  - Clicked on the ad and visited the website.
  - Spent more time on the site, viewed multiple pages.
  - Engaged with interactive tools like the "Price Estimator."

- **Non-Interested Prospects (Assumed 70%):** 😕
  - Clicked on the ad but didn't engage much.
  - Bounced quickly from the landing page.

- **Marketing Team's Assumed Clicked Customers:** 📊
  - Total Clicks: 193
  - Interested Prospects (30% of clicks): 58
  - Non-Interested Prospects (70% of clicks): 135

Business Recommendation & Simulation

Expected Revenue Sales and CVR (Conversion Rate): 💰📈

- **Marketing Team's Expected Results:**
  - Total Conversions: 58 (assuming all 58 interested prospects converted).
  - Total Revenue Sales: $11,600 (58 conversions x $200 average transaction value).
  - Conversion Rate (CVR): 100% (since all expected interested prospects are assumed to convert).

- **Non-Interested Prospects:** 💰
  - Assuming a 2% conversion rate for non-interested prospects.
  - Expected Conversions: 3 (135 x 2%).
  - Average Transaction Value: $200
  - Revenue from Non-Interested Prospects: $600 (3 x $200).

- **Total Expected Results:**
  - Total Conversions: 61 (58 from interested + 3 from non-interested)
  - Total Revenue Sales: $12,200 ($11,600 from interested + $600 from non-interested)
  - Conversion Rate (CVR): Approximately 31.6% (61 conversions out of 193 clicks)

For more information about the project, Click Here

Business Recommendation & Simulation

	Not Clicked The Ads	Clicked The Ads
Not Clicked The Ads	88	6
Clicked The Ads	5	94

Certainly, let's explain that the model was created using the **DecisionTreeClassifier** and compare the revenue and Conversion Rate (CVR) with and without the machine learning model

It achieved an impressive accuracy of **94.3%**. The confusion matrix revealed the following:

- **True Positives (TP)**: 94
- **True Negatives (TN)**: 88
- **False Positives (FP)**: 6
- **False Negatives (FN)**: 5

For more information about the project, Click Here

Business Recommendation & Simulation

Comparison Revenue with Machine Learning and without Machine Learning

- **Marketing Team's Expected Results:**
  - Total Conversions: 58 (assuming all 58 interested prospects converted).
  - Total Revenue Sales: $11,600 (58 conversions x $200 average transaction value).
  - Conversion Rate (CVR): 100% (since all expected interested prospects are assumed to convert).

- **Machine Learning Model's Results:**
  - Total Conversions: 182 (as previously mentioned).
  - Total Revenue Sales: $36,400 (as previously mentioned).
  - Conversion Rate (CVR): Approximately 94.3% (as previously mentioned).

Comparison Summary:
With Machine Learning (ML) 🤖:

The model predicts significantly higher conversions and revenue than the marketing team's initial expectations. ML delivers a Conversion Rate (CVR) of around 94.3%, which is lower than the assumed 100% but still quite high. 📈💰Without Machine Learning (ML) The marketing team's expected results would fall short, achieving only 58 conversions and $11,600 in revenue. The CVR remains high, but it doesn't reach 100%. 🙅‍♂️🤷‍♂️

For more information about the project, Click Here

Thank You