Predict Clicked Ads Customer Classification by using Machine Learning

Created by:
Aziz Prabowo
azizprabowo128@gmail.com
linkedin.com/in/aziz-prabowo

A fresh graduate in Data Science with high interest artificial intelligence, data science, and business analytics. Experienced in data cleaning, exploratory data analysis, visualization, machine learning, and basic deep learning through academic, bootcamps, courses, and personal projects.

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

Project Overview

Rakamin Academy

Background

A company in Indonesia wants to evaluate the effectiveness of an advertisement they have launched. This is important for the company to understand how well the advertisement reaches its audience and how effectively it attracts customers to view the ad. By processing historical advertisement data and uncovering insights and patterns, the company can better determine their marketing targets.

Goal

Analyze historical advertisement data, engineer relevant features, and evaluate multiple models to uncover key factors influencing customer interaction, ultimately providing insights to optimize marketing strategies.

Objective

Develop a machine learning classification model to accurately identify target customers who are likely to engage with advertisements, improving the effectiveness of marketing campaigns.

Link to Jupyter Notebook

Dataset Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   daily_time_spent_on_site  987 non-null    float64
 1   age                       1000 non-null   int64
 2   area_income               987 non-null    float64
 3   daily_internet_usage      989 non-null    float64
 4   gender                    997 non-null    object
 5   timestamp                 1000 non-null   object
 6   clicked_on_ad             1000 non-null   object
 7   city                      1000 non-null   object
 8   province                  1000 non-null   object
 9   category                  1000 non-null   object
dtypes: float64(3), int64(1), object(6)
memory usage: 78.2+ KB
```

This dataset contains **user behavior data related to online advertising**, with **1,000 total entries and 10 columns** detailing various user attributes.

- **User Demographics:** Includes features such as **Age**, **Gender**, **City**, and **Province**, along with **Area_Income** that reflects average income in the user's region.

- **Online Behavior:** Variables like **Daily_Time_Spent_On_Site** and **Daily_Internet_Usage** capture user activity and engagement with the internet and specific websites.

- **Advertising Interaction:** The key target variable is **Clicked_On_Ad** (binary: clicked / not clicked), complemented by **Category** of the ad and **Timestamp** indicating when the ad was shown.

Data Cleaning

```
[3]: df.isnull().sum()

[3]: daily_time_spent_on_site    13
     age                          0
     area_income                 13
     daily_internet_usage        11
     gender                       3
     timestamp                    0
     clicked_on_ad                0
     city                         0
     province                     0
     category                     0
     dtype: int64
```

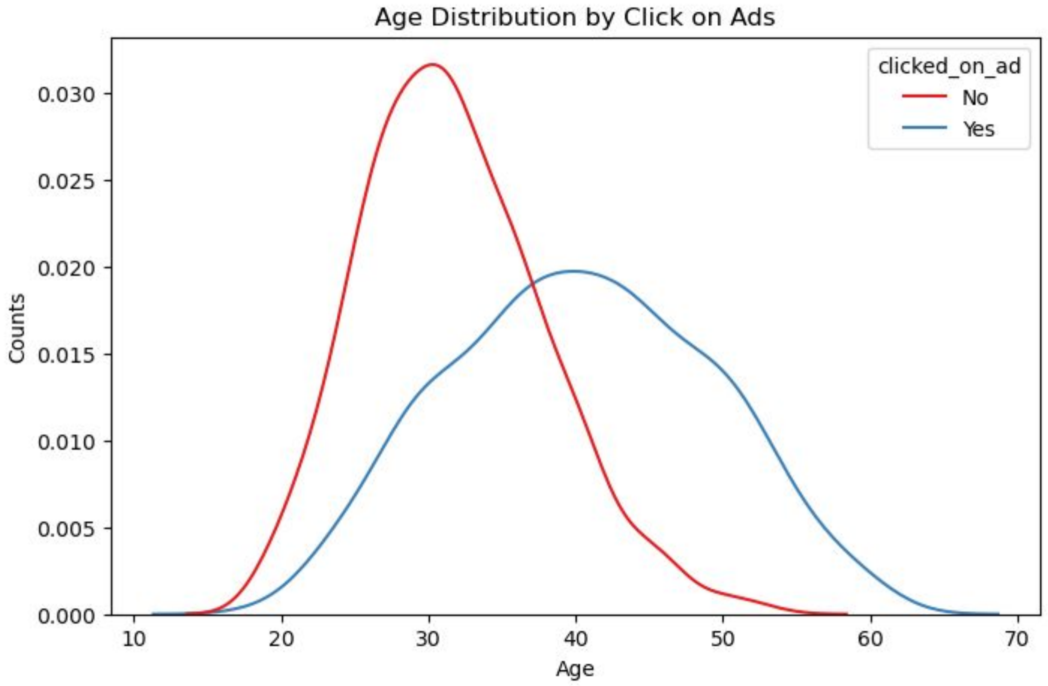
Statistical Imputation

```
[5]: df.isnull().sum()

[5]: daily_time_spent_on_site    0
     age                         0
     area_income                 0
     daily_internet_usage        0
     gender                      0
     timestamp                   0
     clicked_on_ad               0
     city                        0
     province                    0
     category                    0
     dtype: int64
```

Missing values in the `daily_time_spent_on_site` and `daily_internet_usage` columns were imputed using the median because their distributions are skewed, making the median more suitable to avoid the influence of outliers. The `area_income` column was imputed with the mean because its distribution is more normal, while for the `gender` column, which is categorical data, missing values were imputed using the mode to reflect the most frequently occurring value.

Exploratory Data Analysis

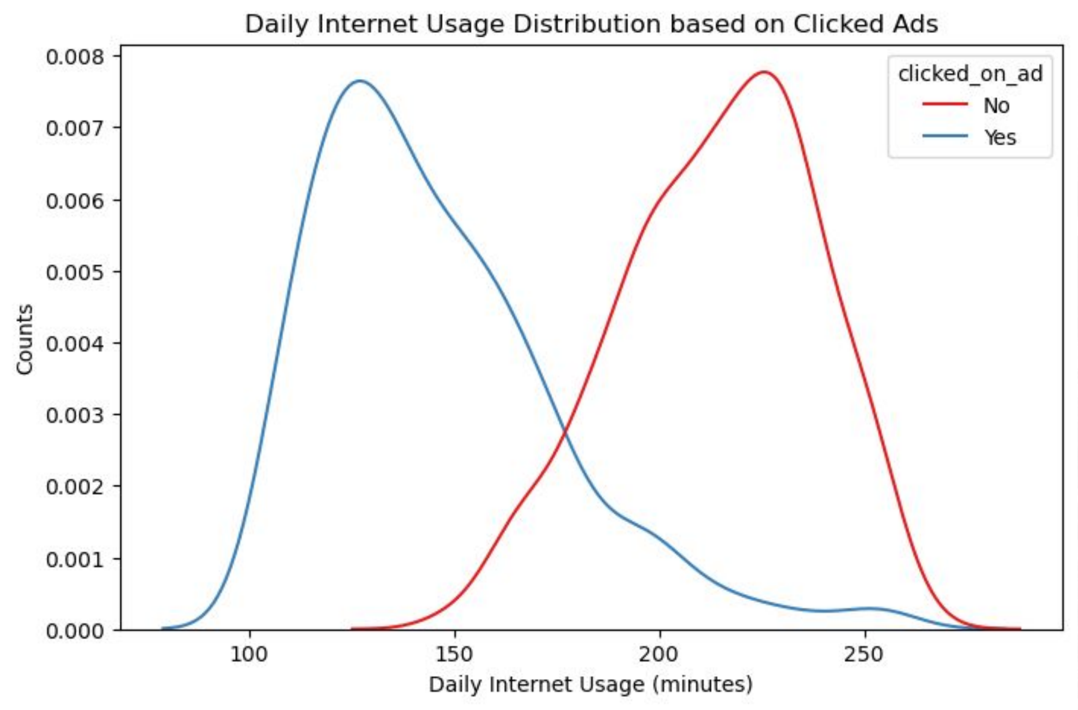
Age in the 40s More Responsive to Ads



Customers who did not click on the ad are predominantly in the 28–36 age group, while those who clicked tend to be older, peaking around 40 years old. **The more mature target market, especially those in their 40s, shows a higher response to the ads**, whereas the 30s age group tends to be less interested.

Exploratory Data Analysis

Internet Usage Duration Influences Ad Clicks

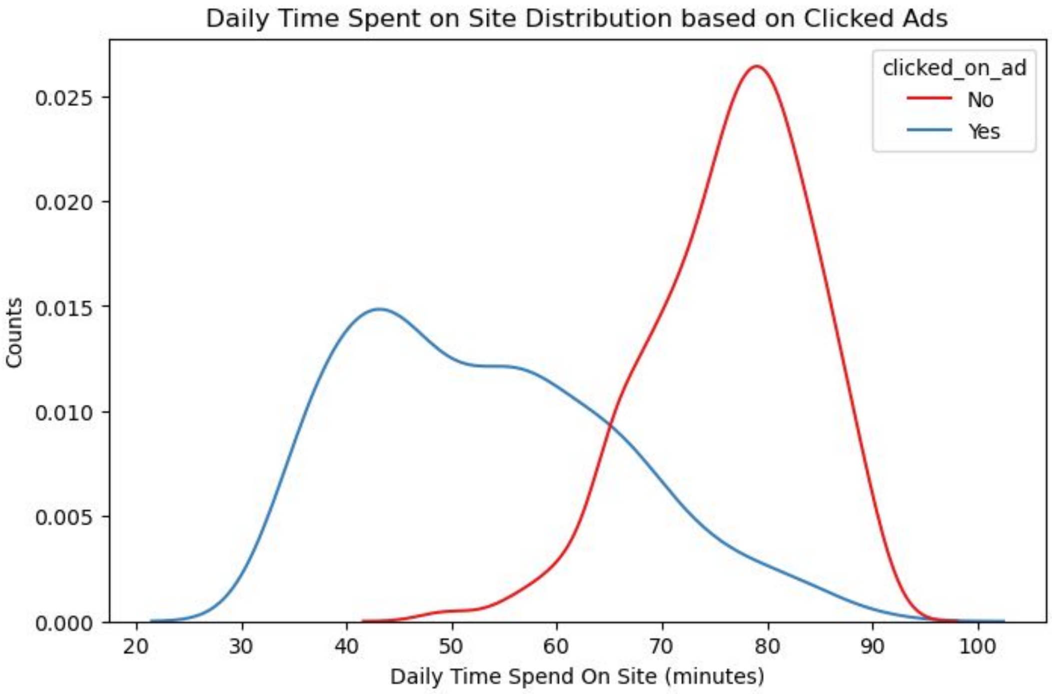


Customers who clicked on the ad generally have a daily internet usage duration between 120 and 160 minutes, which is under 3 hours. In contrast, customers who did not click on the ad tend to have longer internet usage durations, typically ranging from 200 to 230 minutes.

Exploratory Data Analysis

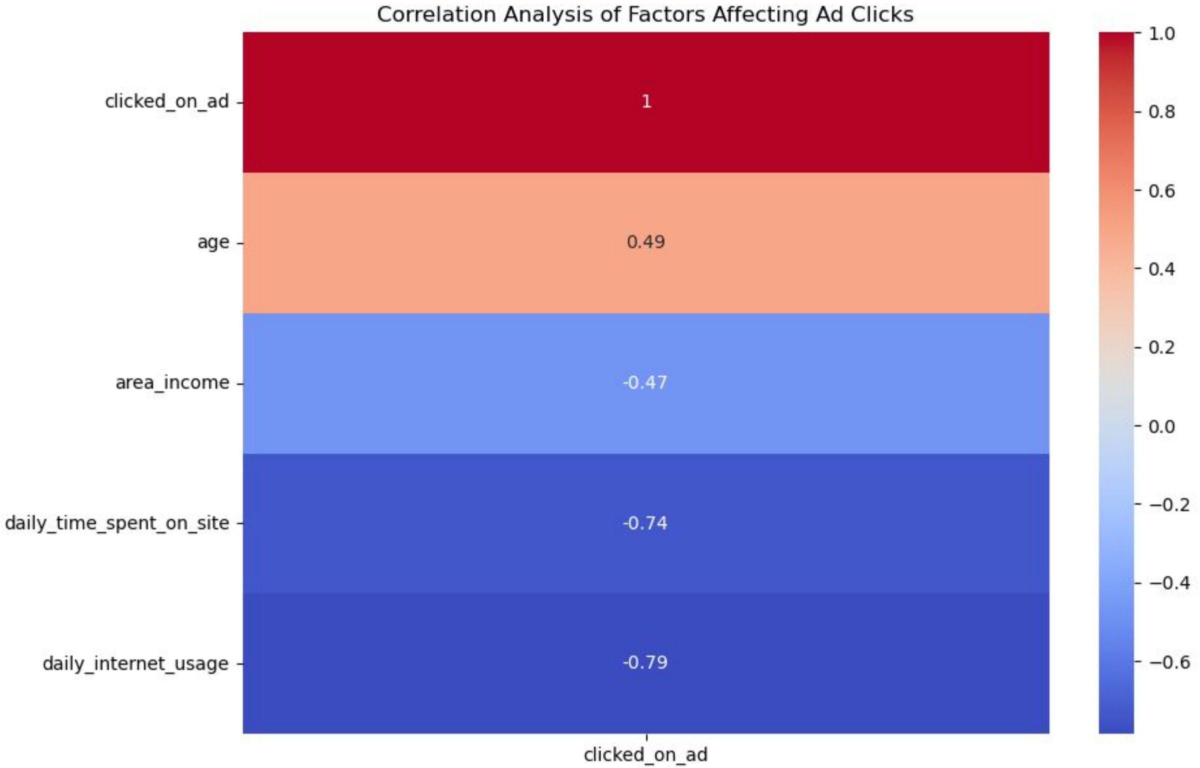
Rakamin Academy

Time Spent on Website Affects Ad Clicks



Customers who clicked on the ad generally spend less than 1 hour on the website, with a wider time range of 42–60 minutes. In contrast, customers who did not click tend to spend more than 1 hour, but within a narrower range of 72–82 minutes.

Exploratory Data Analysis



There is a strong negative correlation between "clicked_on_ad" and both "daily_time_spent_on_site" (-0.74) and "daily_internet_usage" (-0.79), meaning that **the more time users spend on the site or the internet, the less likely they are to click on ads**. On the other hand, there is a moderate positive correlation between "clicked_on_ad" and "age" (0.49), indicating that **older users are more likely to click on ads**. Additionally, there is a moderate negative correlation between "clicked_on_ad" and "area_income" (-0.47), suggesting that **users with higher income tend to click less on ads**.

Feature Engineering

Rakamin Academy

New Features Generated from timestamp

	hour	day_of_week	is_weekend	part_of_day
662	22	5	True	night
426	9	1	False	morning
500	13	3	False	afternoon
340	0	0	False	night
572	23	6	True	night

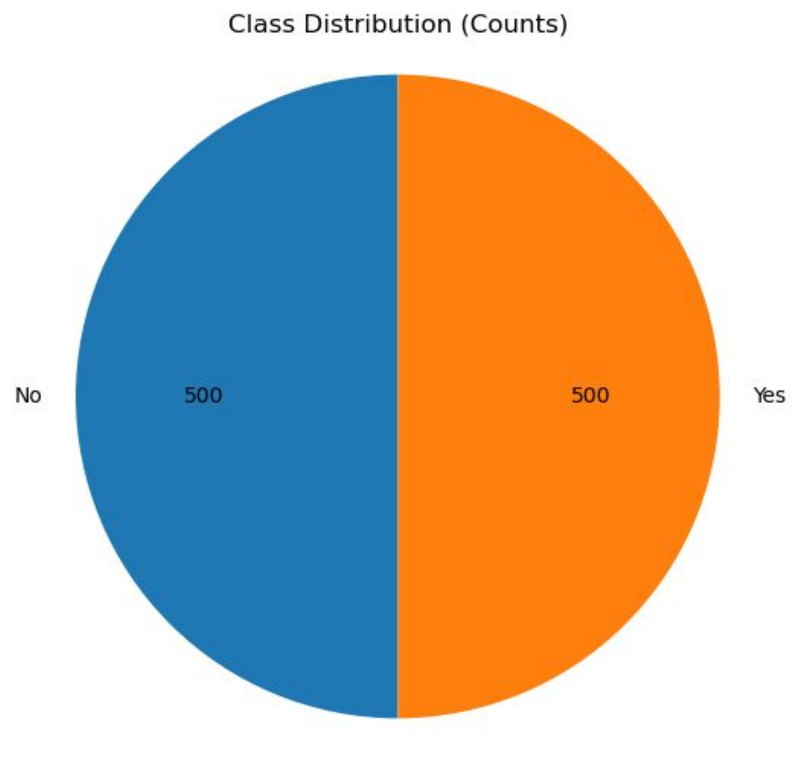
As a result of the feature engineering process, several new features have been added to the dataset to enrich the available information. These features include:

- **hour**: indicates the hour when the user accessed the site.
- **day_of_week**: shows the day of the week when the access occurred (with 0 = Monday through 6 = Sunday).
- **is_weekend**: an indicator of whether the access happened on a weekend (Saturday or Sunday).
- **part_of_day**: groups the access time into parts of the day, namely morning, afternoon, evening, or night.

Data Preprocessing

Rakamin Academy

Class Distribution is Balanced



The class distribution for this dataset is balanced, with 500 non-clicks and 500 clicks out of a total of 1,000 rows. Therefore, **there is no need to perform any class imbalance handling procedures**.

Data Preprocessing

Sample Categorical Features after Encoding

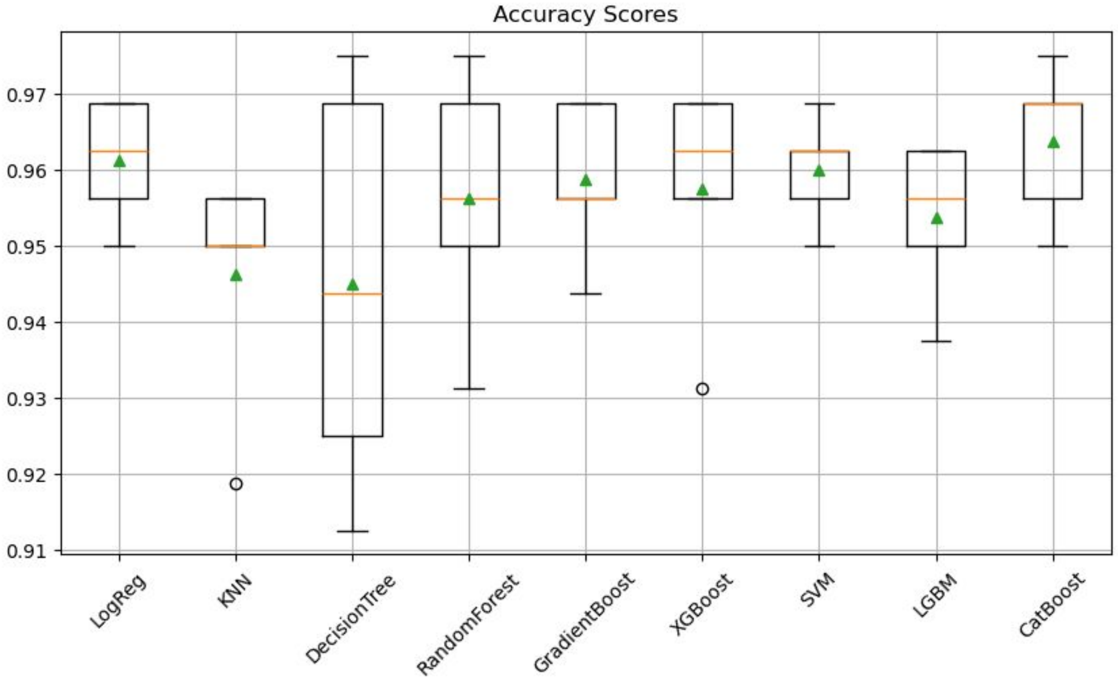
```
     clicked_on_ad  hour  is_weekend  province_Bali  province_Banten  ...  \
719            Yes     5           1          False            False  ...
426            Yes     9           0           True            False  ...
538             No     0           1          False            False  ...
423            Yes     9           0          False             True  ...
645            Yes    13           0          False            False  ...

     day_of_week_1  day_of_week_2  day_of_week_3  day_of_week_4  \
719          False          False          False          False
426           True          False          False          False
538          False          False          False          False
423          False           True          False          False
645          False          False          False          False

     day_of_week_5  day_of_week_6  part_of_day_afternoon  part_of_day_evening  \
719           True          False                  False                False
426          False          False                  False                False
538           True          False                  False                False
423          False          False                  False                False
645          False          False                   True                False

     part_of_day_morning  part_of_day_night
719                 True              False
426                 True              False
538                False               True
423                 True              False
645                False              False

[5 rows x 75 columns]
```

The categorical columns in the dataset have been successfully converted into numeric format using one-hot encoding with the pd.get_dummies() function. Initially, there were 5 categorical columns in the dataset. After the encoding process, these five columns generated a total of **67 new columns**—one for each unique category. The total number of columns in the dataset increased from **13 before encoding** to **75 after encoding**.

Data Modeling

Rakamin Academy

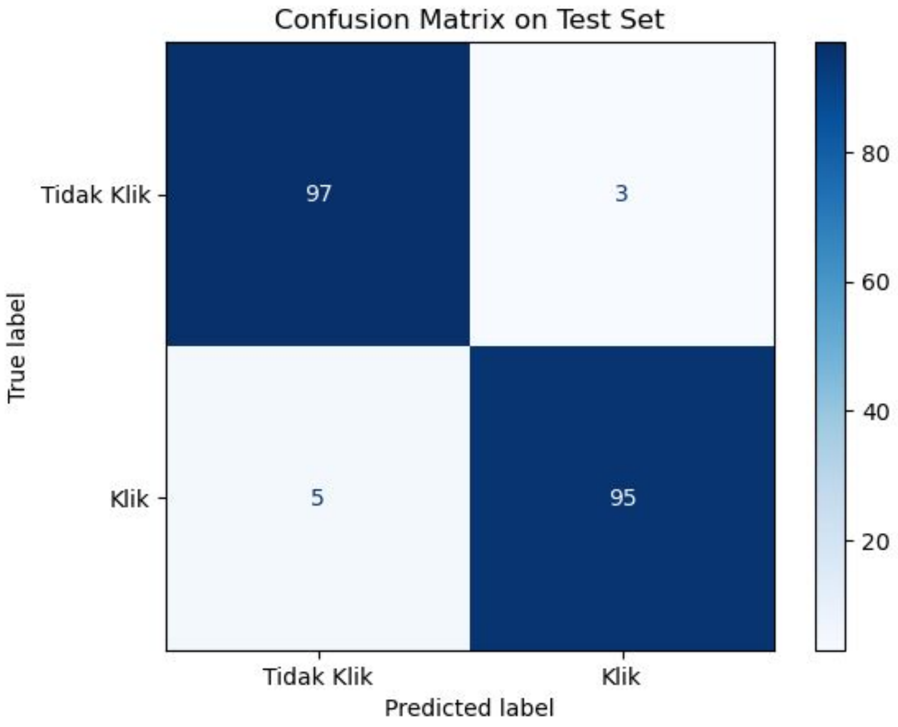
Models Training Comparison



Overall, most models perform at a high accuracy level (above 0.95), with **CatBoost, Logistic Regression, and XGBoost demonstrating the most stable and consistently high results.** Random Forest, Gradient Boost, and SVM also perform competitively with relatively tight score distributions. In contrast, Decision Tree and KNN show more variability and occasional lower accuracy, making them less reliable compared to ensemble methods and advanced boosting models. This indicates that ensemble and boosting techniques (CatBoost, XGBoost, Gradient Boost) are generally more robust and reliable choices for achieving high accuracy in this task. Among all the models tested, **CatBoost achieved the highest accuracy with a score of Accuracy: 0.964**.

Data Modeling

CatBoost Confusion Matrix



On the test set containing 200 data points, the CatBoost model successfully predicted **95 out of 100 customers who clicked on the ad** (TP = 95, FN = 5), and made only **3 false positive predictions** by incorrectly classifying non-clickers as clickers (FP = 3).

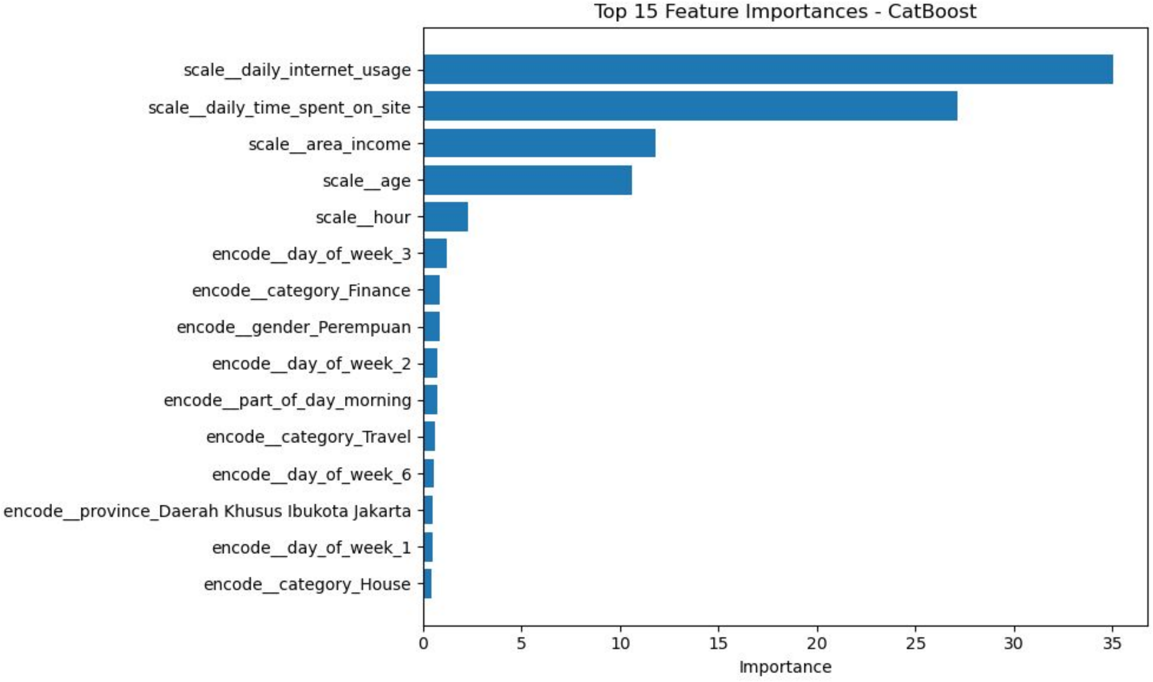
The **recall** is **0.95** for the *click* class and **0.97** for the *non-click* class. The **precision** is also high: **97% for clicks** and **95% for non-clicks**.

Overall, the model achieved an **accuracy of 96%**, meaning that 96 out of every 100 predictions made by the model were correct.

Data Modeling

Rakamin Academy

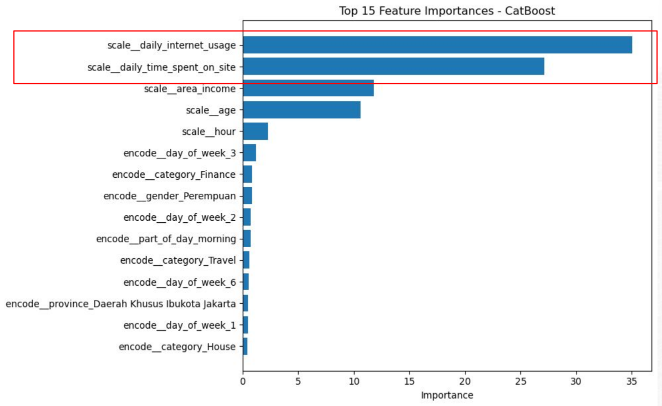
Feature Importance

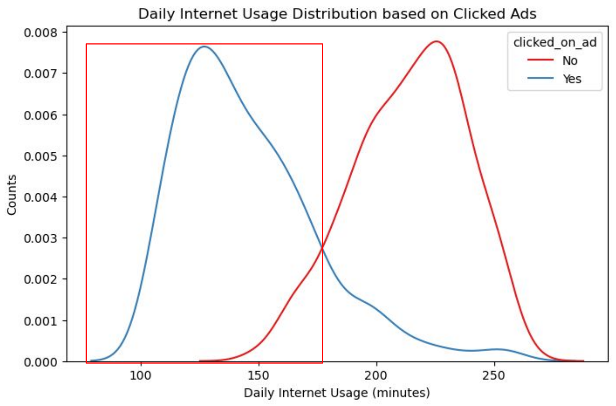


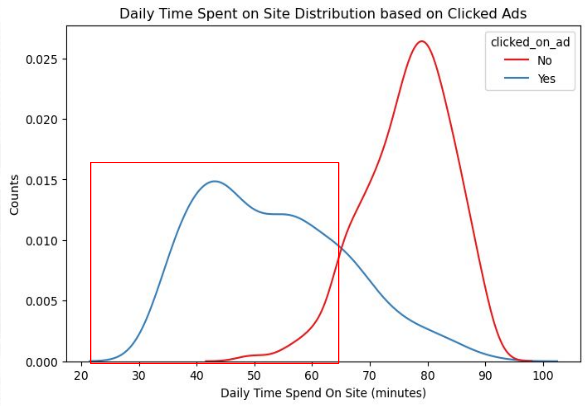
The feature importance analysis from this experiment shows that the two features contributing most to predicting whether a customer will click on an ad remain the same as in the previous experiment. This indicates that **the intensity of internet usage and the amount of time users spend on the website are key indicators of interest in the ad**. In other words, the more frequently someone uses the internet and the longer they stay on the site, the more likely they are to be interested in and click on the displayed ad.

Business Recommendation & Simulation

Two Most Influential Features: Daily Internet Usage and Time on Site

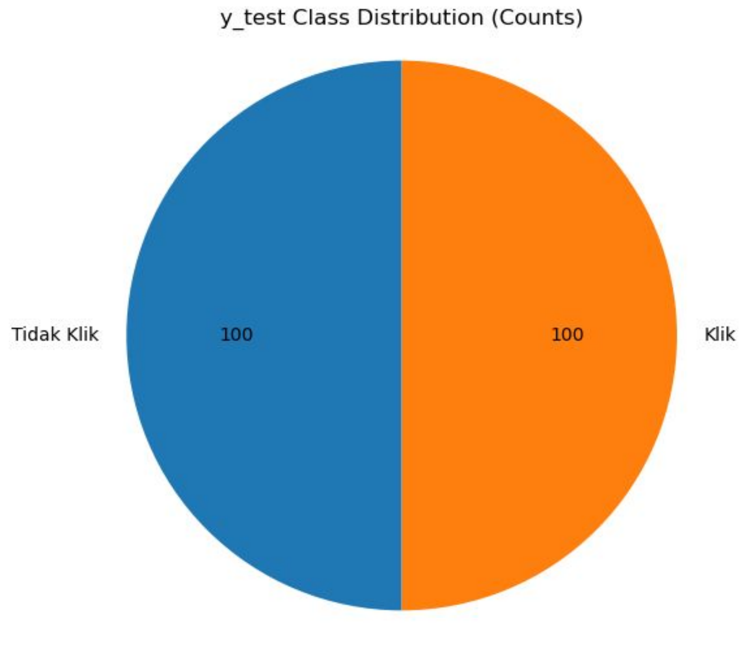






Feature importance results indicate that **daily_internet_usage** and **daily_time_spent_on_site** are the two most decisive features in predicting ad clicks. This finding is supported by the EDA: users with **lower site visit duration** and **less internet usage** tend to be **more likely to click on ads**.

Business Recommendation & Simulation



Scenario 1: Without Using Machine Learning (Conventional Strategy)

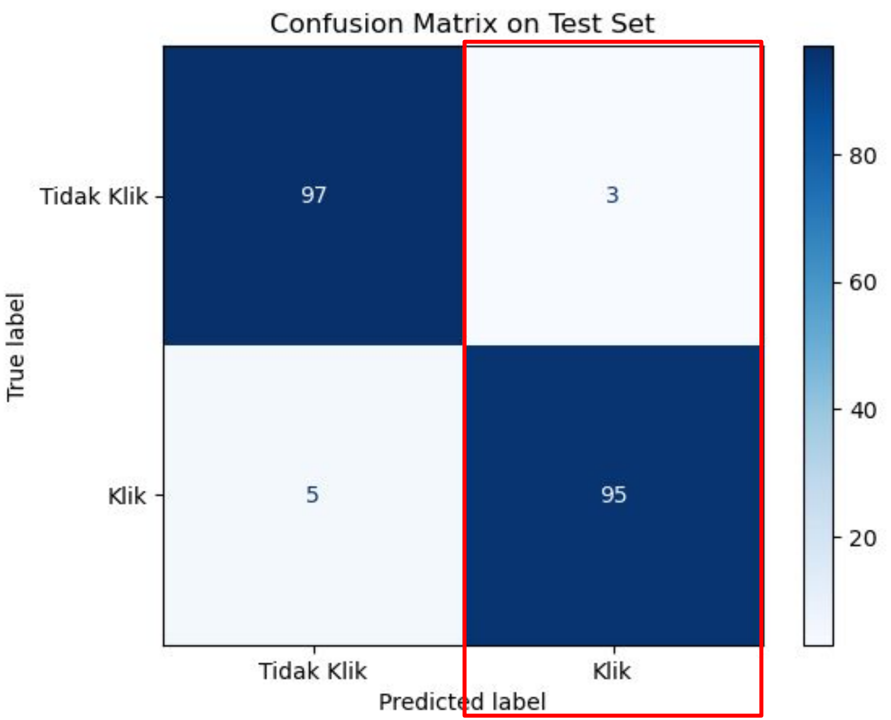
In this scenario, the marketing campaign targets 200 users (test data) randomly. The average conversion rate (ad click rate) is 50%, based on the balanced dataset.

- **Number of Users**: 200
- **Marketing Cost per User**: Rp 10,000
- **Total Cost**: 200 × Rp 10,000 = **Rp 2,000,000**
- **Conversion Rate**: 50%
- **Number of Ad Clicks (Target Achieved)**: 200 × 50% = **100 clicks**
- **Revenue per Click (assumption)**: Rp 15,000
- **Total Revenue**: 100 × Rp 15,000 = **Rp 1,500,000**
- **Profit**: Rp 1,500,000 − Rp 2,000,000 = **−Rp 500,000**

Without using a machine learning model, the company would incur a **loss of Rp 500,000**.

Business Recommendation & Simulation

Rakamin Academy



Scenario 2: Using Machine Learning (Data-Driven Strategy)

In this scenario, the company uses a machine learning model to identify 200 users most likely to click on the ad. The model achieves a **precision of 97%** and a **recall of 95%** for the *click* class on the test set.

- **Number of Users**: 200
- **Marketing Cost per User**: Rp 10,000
- **Total Cost**: 200 × Rp 10,000 = **Rp 2,000,000**
- **Conversion Rate**: 97% (based on the model's precision for click predictions)
- **Number of Ad Clicks (Target Achieved)**: 200 × 97% = **194 clicks**
- **Revenue per Click (assumption)**: Rp 15,000
- **Total Revenue**: 194 × Rp 15,000 = **Rp 2,910,000**
- **Profit**: Rp 2,910,000 − Rp 2,000,000 = **Rp 910,000**

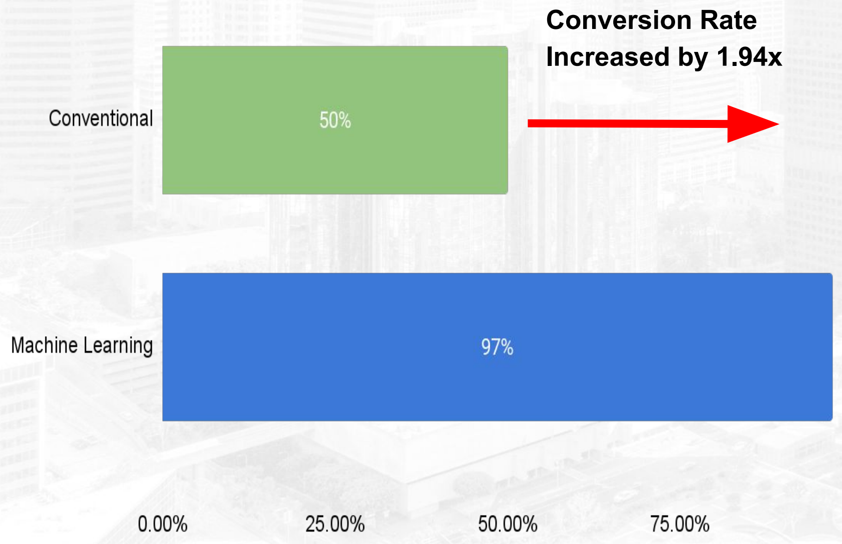
By leveraging a machine learning model, the company would generate a **profit of Rp 910,000**—a clear improvement over the conventional approach.

Business Recommendation & Simulation

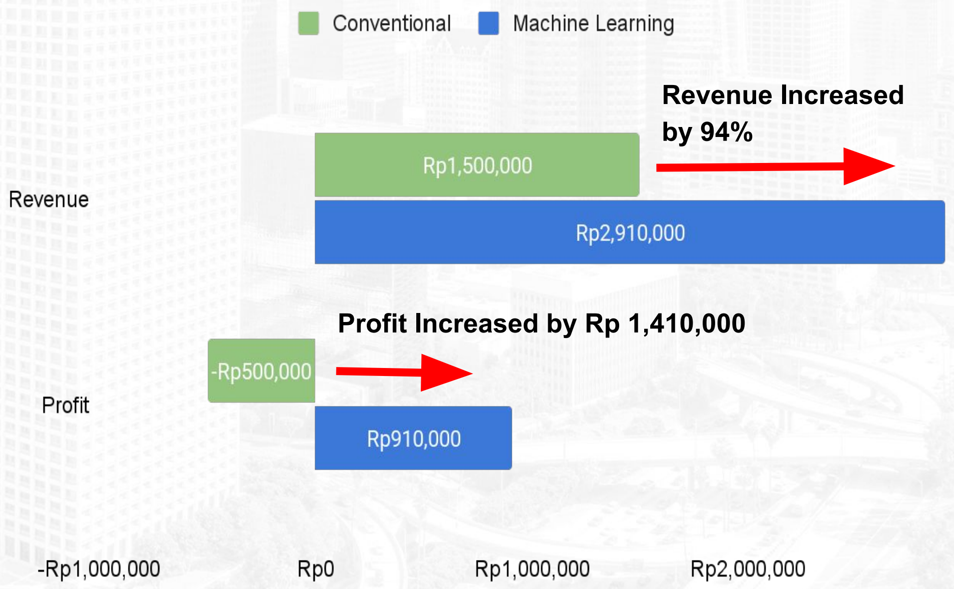
Rakamin Academy

Conventional Strategy vs Machine Learning

Conversion Rate



Revenue & Profit



Business Recommendation & Simulation

Rakamin Academy

Conclusion

This simulation demonstrates that using a machine learning model for marketing targeting significantly improves the conversion rate—from **50% to 97%**. This increase leads to an additional profit of **Rp 1,410,000**, shifting from a **loss of Rp 500,000** to a **profit of Rp 910,000**, all with the **same marketing budget**.

Recommendations

- **Target the Most Responsive Audience**
  Focus ad campaigns on users who spend **less than 60 minutes per day** on the site and use the internet for **less than 180 minutes**. This group has shown the highest likelihood of clicking on ads.
- **Prioritize High-Performing Cities**
  Maximize campaign impact by concentrating on cities with the highest ad performance, such as **South Jakarta, Central Jakarta, and Semarang**.
- **Allocate Budget to High-Impact Product Categories**
  Increase ad spend for product categories with **high feature importance**—such as **furniture**—which show a strong correlation with ad clicks.

Potential Impact

- **Conversion Rate Improvement**: From 50% to 97%, nearly doubling the effectiveness of the campaign.
- **Profit Increase**: Rp 1.41 million improvement, highlighting the financial benefit of a data-driven strategy.
- **Cost Efficiency**: Potential to reduce Customer Acquisition Cost (CAC) by up to **30%** in high-performing cities.
- **CTR Growth**: Overall **Click-Through Rate** (CTR) can increase by **15%–20%** by focusing on the most responsive user segments.

Thank You!