Rakamin Academy

Analysis & Prediction of Online Ad Click Behavior

Mini Project -
Rakamin Academy

By:
Az-Zukhrufu Fi Silmi Suwondo

Rakamin Academy

Problem & Objectives

Problem

Companies must understand the **key factors that influence the effectiveness of online advertising** in order to **improve click-through rates** and **optimize digital marketing strategies** based on user demographics and behavioral patterns.

Goal

Develop and implement a **machine learning model** to **predict click probability** in online advertising with a **minimum precision of 90%**, aiming to increase campaign **CTR from 50% to 65%** (a 30% improvement) and **reduce CPC by 20%** within a **three-month period** from January to March 2025.

Objectives

- Analyze the **impact of demographic variables** and **online behavior** on **ad responsiveness**.
- Develop a **high-performance classification model** to **predict 'Clicked on Ad'** outcomes.

Success Metrics

+30%

Click Through Rate (CTR)

More users are clicking on ads after optimization

-20%

Cost per Click (CPC)

Lower cost per click through improved targeting

Dataset Overview

Rakamin Academy

Dataset Summary

- Total Records: 1,000 rows
- Total Features: 16 columns
- Data Period: January 1, 2016 – July 24, 2016

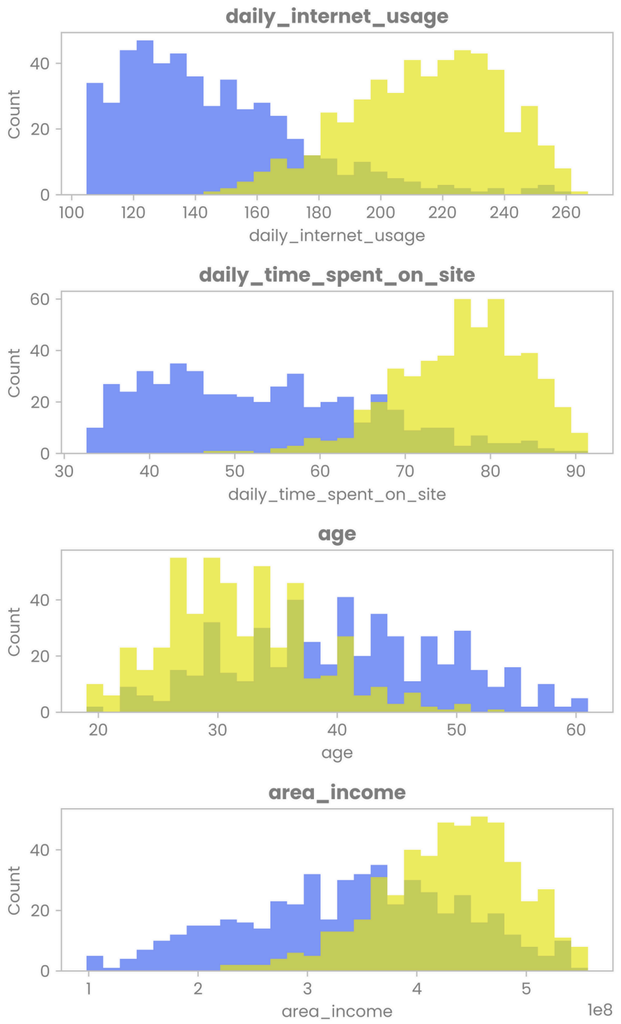
Features

- Daily Time Spent on Site (minutes)
- Age (years)
- Area Income (average income in user's area)
- Daily Internet Usage (minutes)
- Gender (Male)
- Location (City, Province)
- Product Category
- Clicked on Ad (Yes / No)
- Timestamp (interaction time)

Source Data

Rakamin Academy

Profiling Ad Click Behavior



Non-Clickers

Daily internet usage
216 minutes

Daily time spent on site
77.5 minutes

Age
31 years old

Area income
435 million

Ad Clickers

Daily internet usage
139 minutes

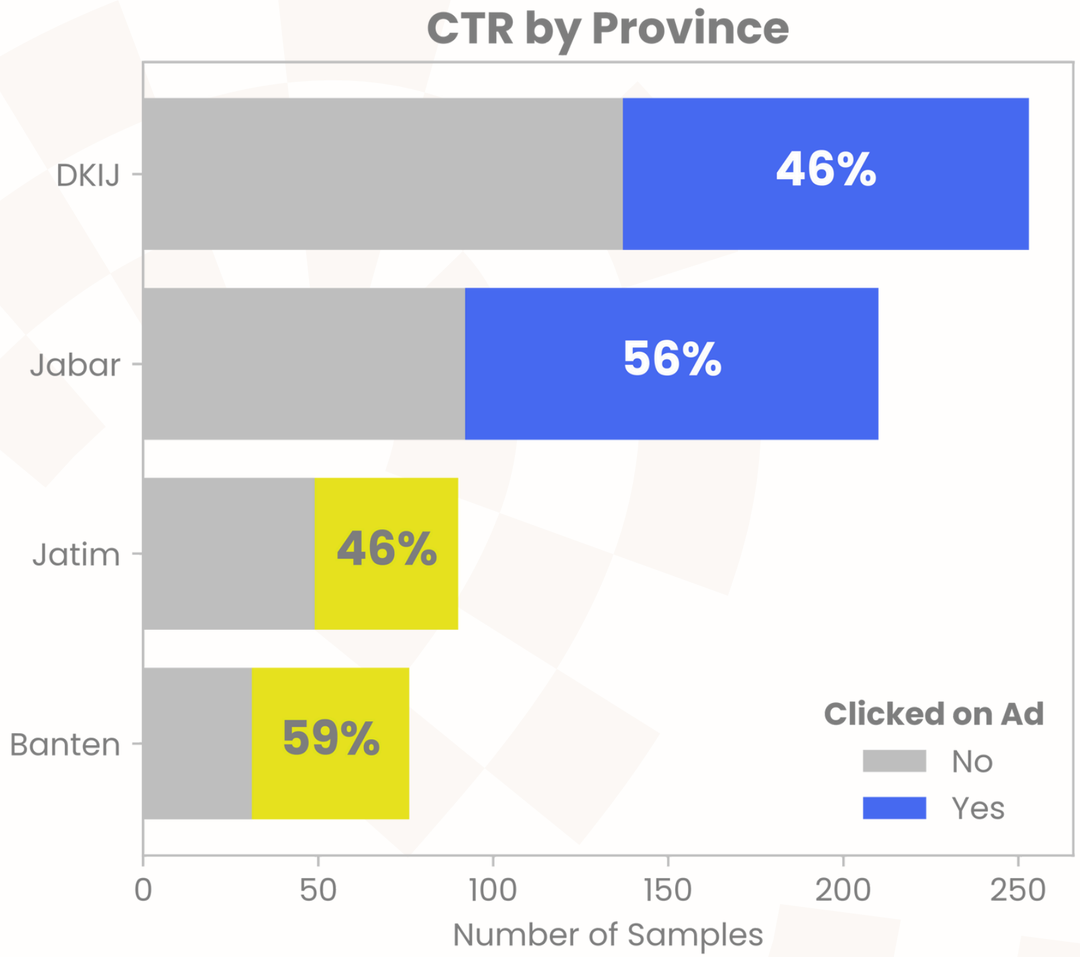
Daily time spent on site
51.6 minutes

Age
40 years old

Area income
347 million

Rakamin Academy

Budget Allocation Optimization



The chart highlights the top **4 provinces** with the **largest audiences**. **West Java** and **Jakarta** stand out with high volume and strong CTR, making them the main focus. **East Java** shows growth potential, while **Banten** is experimental with the highest CTR despite low volume.

Actionable Item

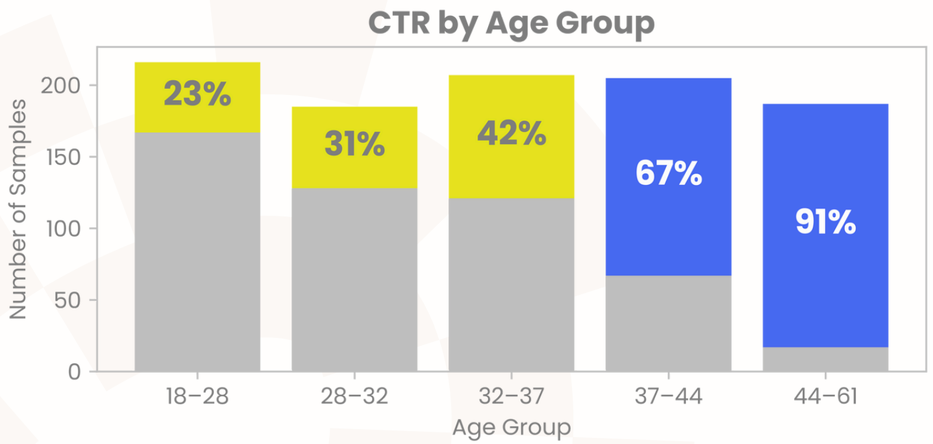
Primary Focus

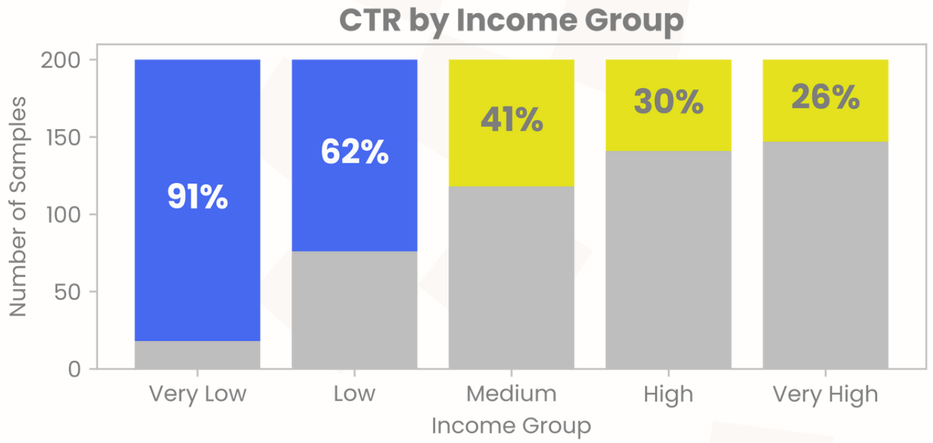
- **Jawa Barat & DKI Jakarta**

High-volume, high-performing channels. Core investment goes here to maintain consistent results and drive reliable ROI.

Rakamin Academy

Audience Targeting Adjustments





CTR is significantly **higher** among **older audiences** and **lower-income groups**, while younger audiences and higher-income segments underperform and present room for improvement. These patterns show that both age and income are **strongly linked to CTR**, making them key factors to focus on for adjustments.

Actionable Item

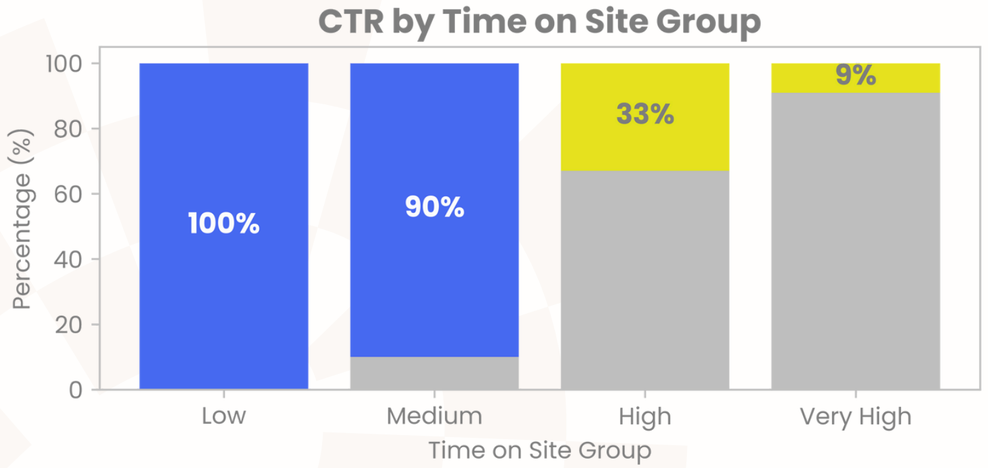
Maintain & Optimize

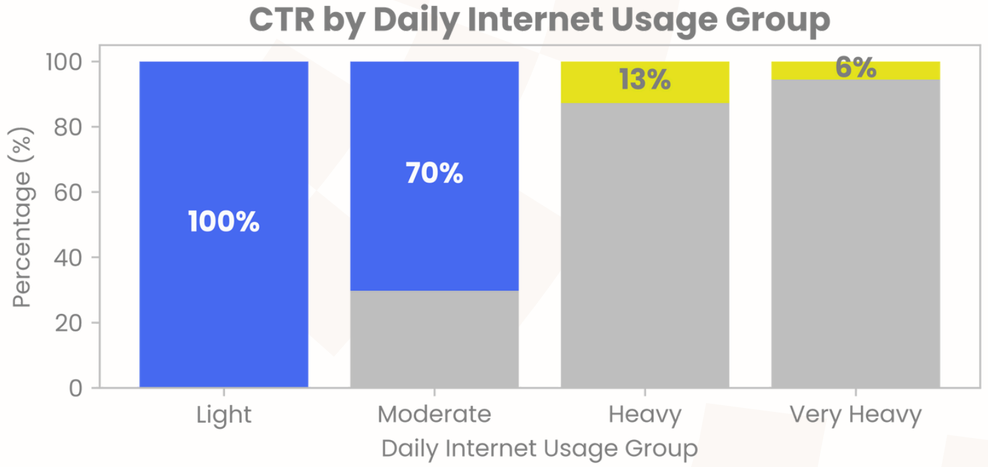
- **Age 46-55**
- **Age 36-45**
- **Low income segment**
- **Low-medium income**

Strategy: Maintain momentum with emotional messaging, practical benefits, and relatable testimonials.

Rakamin Academy

Behavioral Insights & Action Plan





The charts show that **fast site visitors** and **light internet users** deliver the **highest CTR** (up to 100%), indicating strong intent and quick decision-making. In contrast, **CTR drops** sharply among **heavy users** and **long-time visitors**, suggesting they require nurturing strategies instead of direct conversion pushes.

Actionable Item

Target High-Intent Segments

Prioritize light internet users and **fast site visitors**. These users show clear purchase intent and deliver 100% conversion.

Rakamin Academy

Data Preprocessing

Dataset Splitting

- **70%** training
- **30%** testing
- **Stratified by target** variable to preserve class distribution

Missing Values Handling

- **Numerical** columns filled with **median** values from training set
- **Categorical** columns filled with **mode** values from training set

Outliers Handling

- Applied **IQR clipping** to **numerical** features
- Outliers outside were **clipped**

Features Encoding

- **Frequency Encoding**: city, province, category
- **Ordinal Encoding**: age_group
- **One-Hot Encoding**: day_of_week

Rakamin Academy

Data Preprocessing

Feature Scaling

- Applied **StandardScaler** to all numerical features

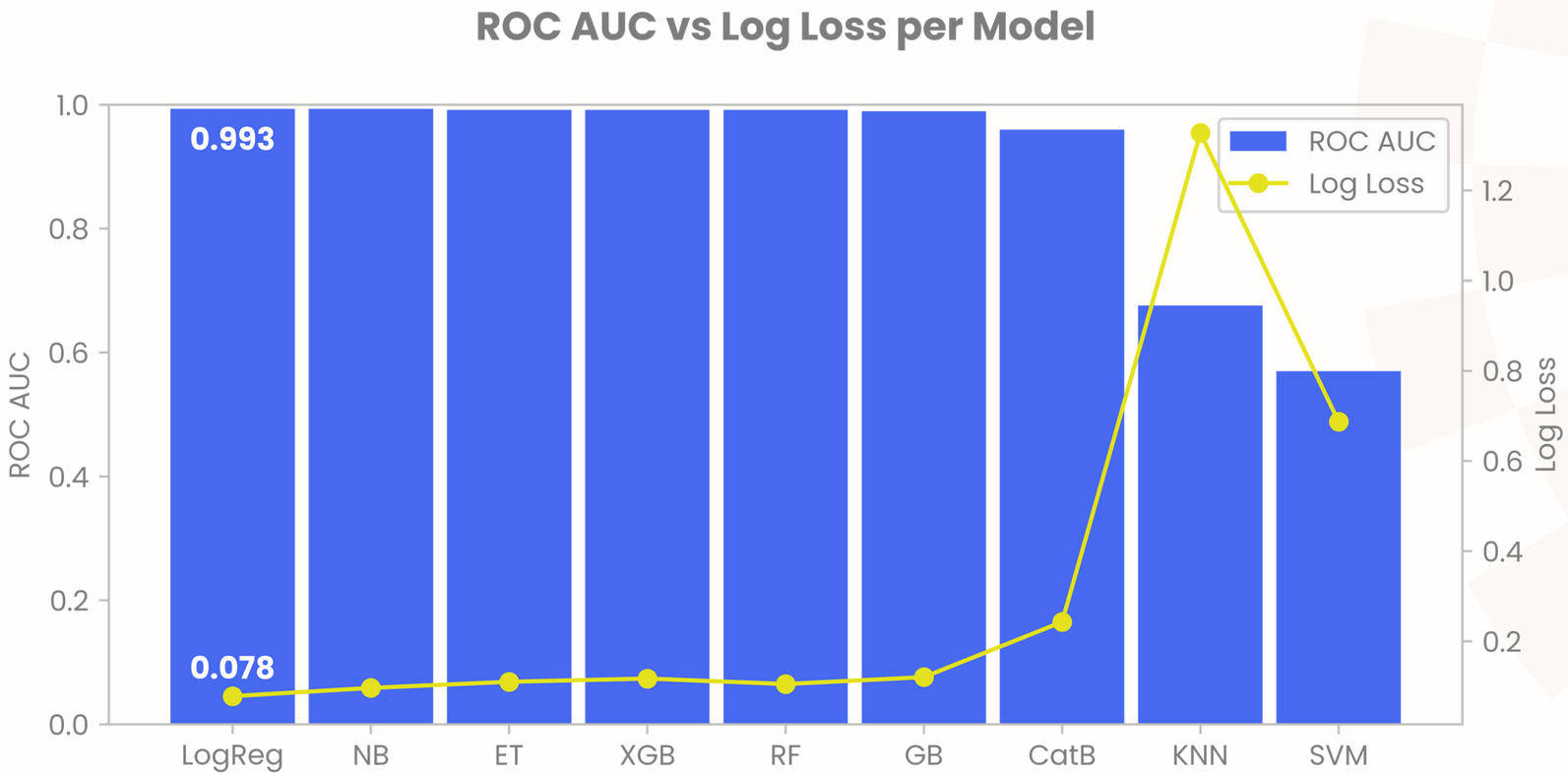
Feature Selection

- Applied Logistic Regression with **L1 regularization** to select the most predictive features

Selected Features

- daily_time_spent_on_site
- age
- area_income
- daily_internet_usage
- male
- time_usage_ratio
- age_group
- hour
- city_freq
- province_freq
- category_freq
- day_of_week_1
- day_of_week_2
- day_of_week_3
- day_of_week_6

Baseline Models



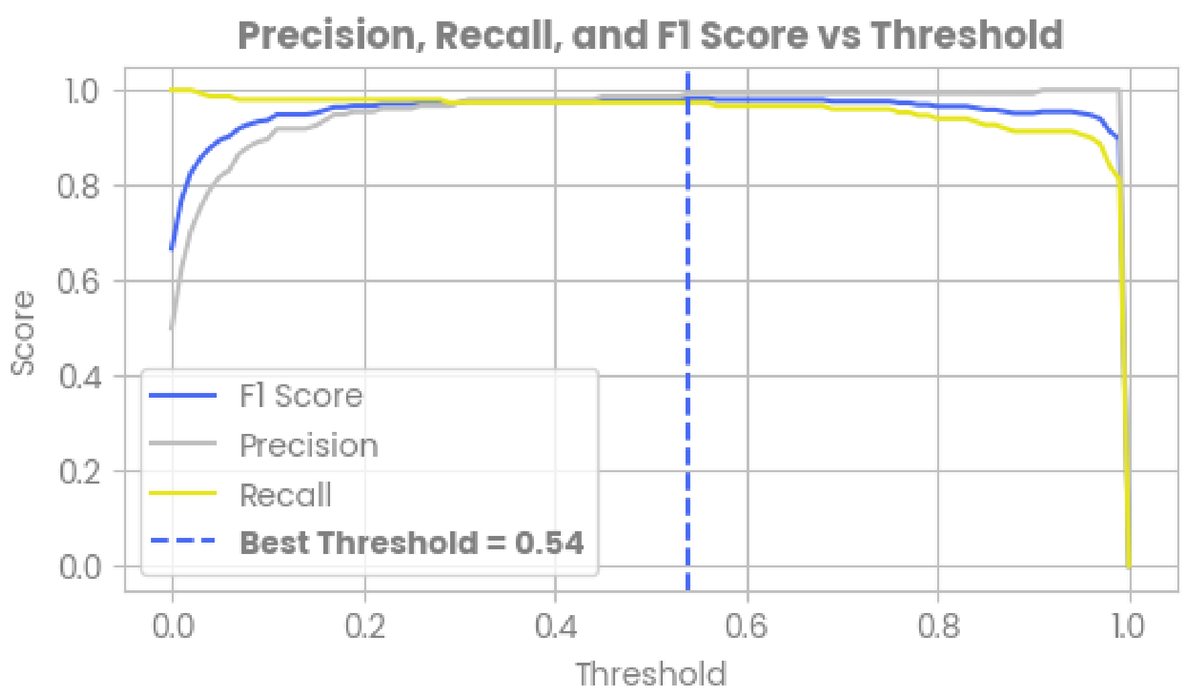
Best Model Selection

Logistic Regression emerged as the top model from nine algorithms, combining **high accuracy** with **strong interpretability**. Using just 15 features, it delivers precise predictions with efficient performance.

- Accuracy: **98.33%**
- AUC: **99.34%**
- F1 Score: **98.32%**
- CV Std Dev: **0.013**

Rakamin Academy

Model Optimization



Model Optimization

Hyperparameter Tuning

- **GridSearchCV** + **Stratified K-Fold** (5 splits)
- Best Parameters:
  - C: 1
  - penalty: 'l2'
  - solver: 'liblinear'
  - max_iter: 500
- Best AUC (Train): **0.9906**

Model Optimization

Threshold Tuning

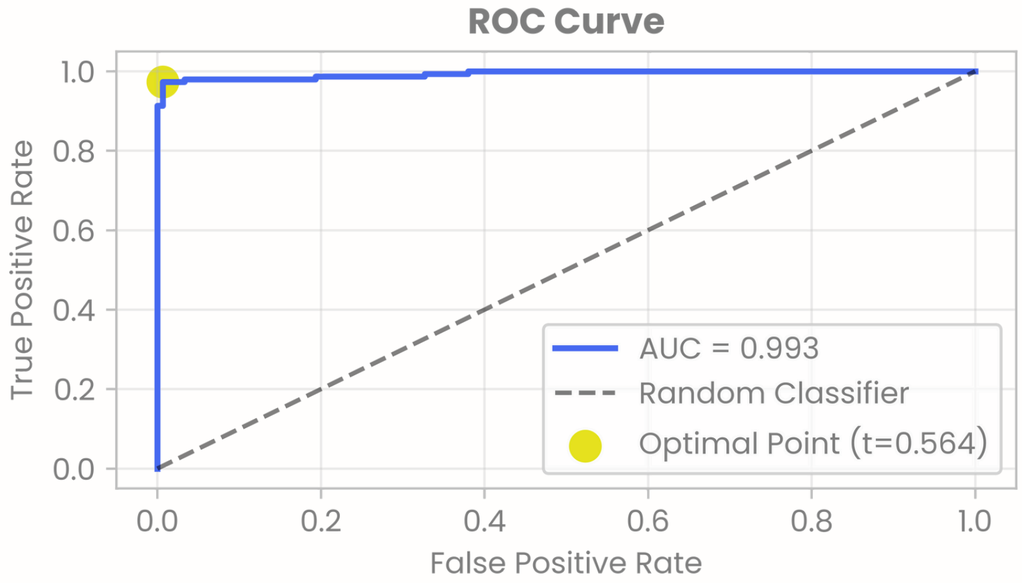
- Optimal Threshold: **0.54**
- Performance Metrics:
  - F1 Score: **0.9832**
  - Precision: **0.9932**
  - Recall: **0.9733**
  - AUC (Test): **0.9934**

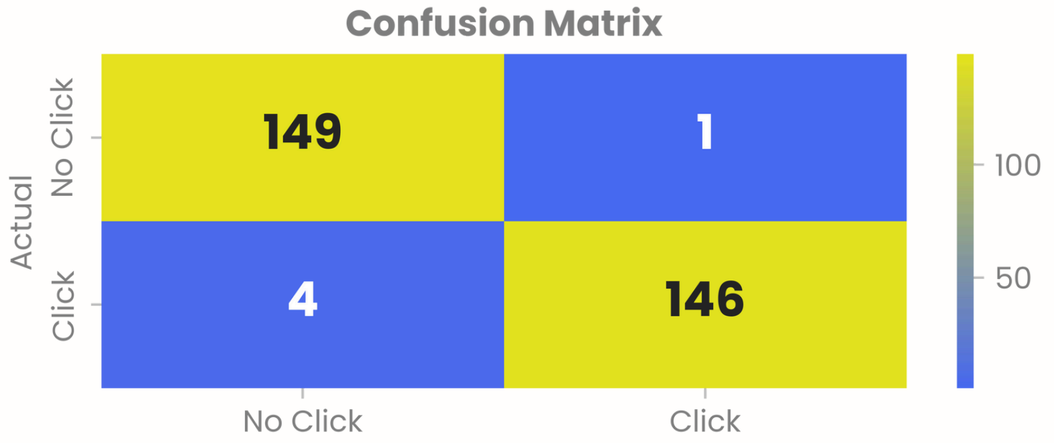
Rakamin Academy

Model Evaluation

- **Precision: 0.9932**
- **Recall: 0.9733**
- **F1-Score: 0.9832**
- **AUC-ROC: 0.9934**

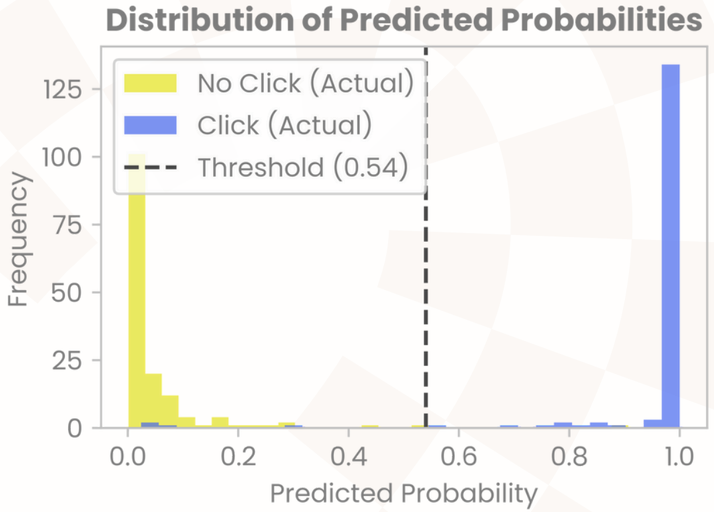
The classification model shows outstanding performance with **near-perfect metrics**. The ROC curve confirms **strong discriminative power**, and only **5 out of 300 test samples were misclassified**. With high accuracy and minimal error, the model is ready for production.

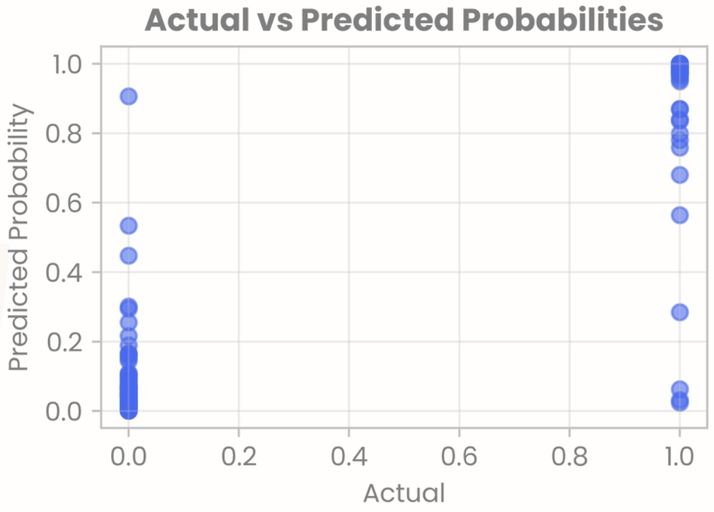


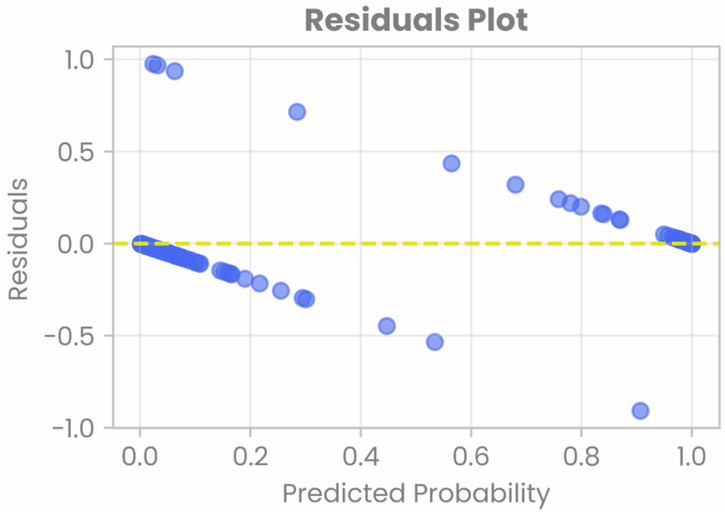


Rakamin Academy

Model Evaluation



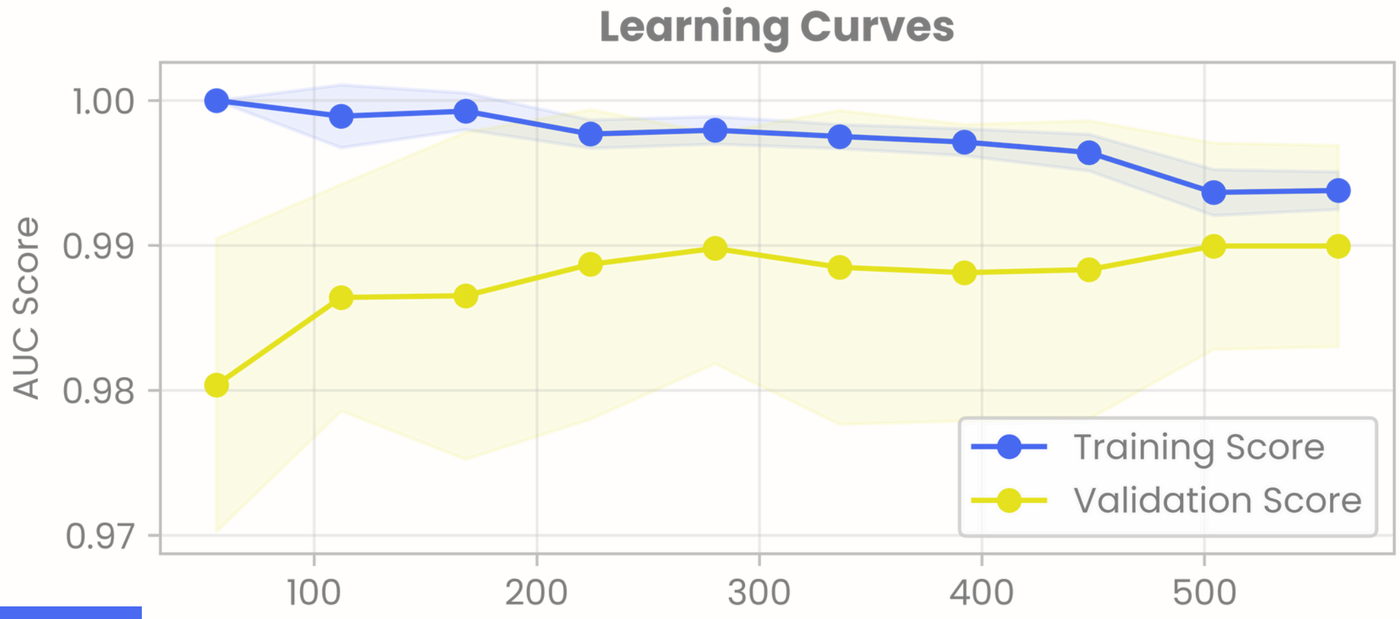




The ad click prediction model **performs well** in **distinguishing between click and no-click users**. Predicted probabilities are mostly near **0 for No Click** and near **1 for Click**, with an **optimal threshold of 0.54**. Actual vs. predicted plots show **good alignment**, and residuals are **small and patternless**, indicating the model is accurate and stable.

Rakamin Academy

Model Stability



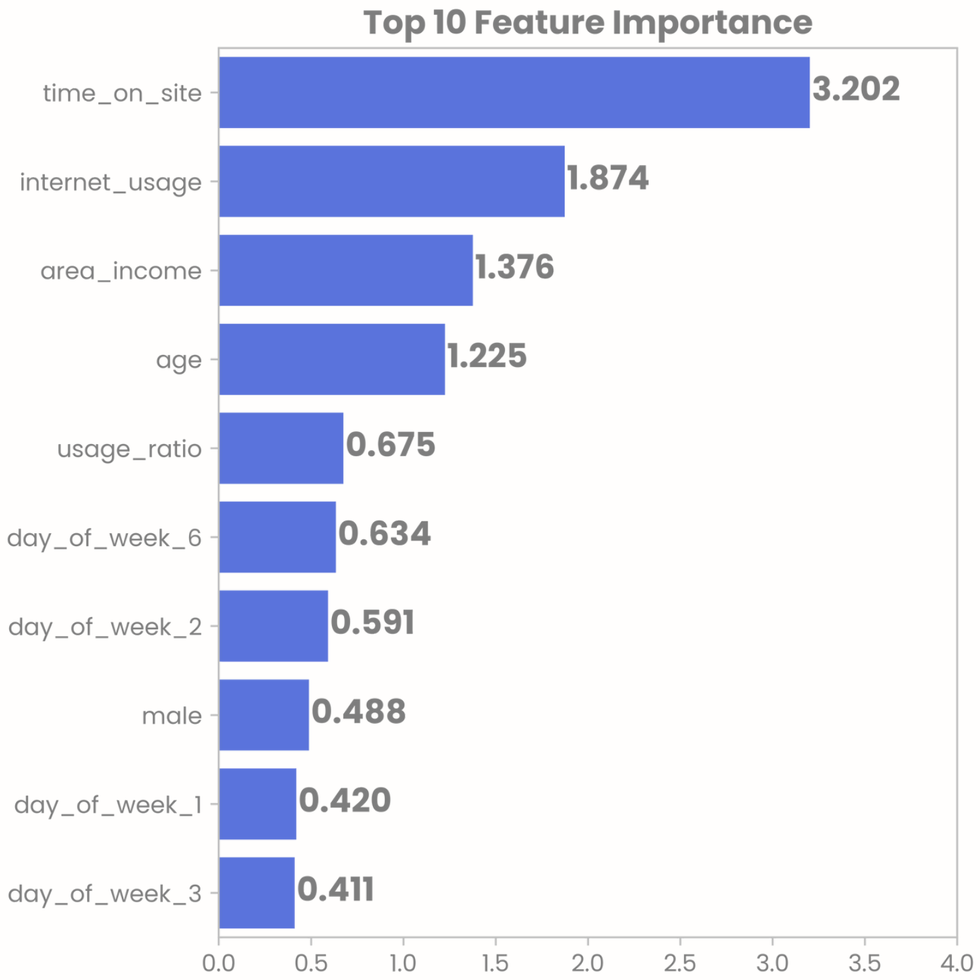
Cross-Validation (5-Fold Stratified)

- ROC-AUC: 0.9900 ± 0.0069 | Gap: 0.0038
- F1-Score: 0.9571 ± 0.0119 | Gap: 0.0127
- Precision: 0.9576 ± 0.0192 | Gap: 0.0185
- Recall: 0.9571 ± 0.0202 | Gap: 0.0064

The learning curves show **strong model stability**, with **training and validation scores converging consistently** across dataset sizes. No signs of overfitting are observed, training remains at ~0.99 while validation steadily improves to 0.989 with minimal gap. Cross-validation confirms **excellent generalization**, with **low variance and small training-validation gaps** (0.38%–1.85%).

Feature Importance

Rakamin Academy



Top Influential Features

- Daily Time Spent on Site
- Age
- Area Income
- Daily Internet Usage

These top four features align with our previous analysis, confirming that targeting based on online behavior, income, and age is the most effective strategy.

Insight

Behavioral Pattern Identified

Users with **low site time** but **high internet usage**, especially in **middle-age** groups from **lower-income** areas, exhibit the highest probability of clicking ads.

Rakamin Academy

Strategic Campaign Summary

Geographic Targeting

- **Primary Focus:** West Java & Jakarta (High volume & strong performance)
- **Experimental Area:** Banten (Highest CTR despite low volume)

Behavioral Intent Signals

- **High-Intent Users:** Short site visits & light internet usage (CTR near 100%)
- **Heavy Users:** Apply nurturing strategy (e.g., retargeting with tailored messaging)

Demographic Prioritization

- **Age Group:** 36–55+ years (Highest CTR)
- **Income Level:** Low to lower-middle (Highest click conversion)

Creative Messaging Strategy

- Use **emotional appeal**
- Highlight **practical benefits**
- Include **relevant testimonials**

Recommendation

Conduct a 1–2 week **A/B test** comparing **current targeting methods (Control Group)** vs. **model-driven targeting (Experiment Group)** to validate these projections in real-world conditions.

Business Impact

99.32%

Click Through Rate (CTR)

Based on model precision of 99.32%, indicating nearly all predicted users are likely to click

-49.6%

Cost per Click (CPC)

Estimated from reduced impressions needed per click, leading to lower cost per acquisition

Executive Summary

Rakamin Academy

This project enhances digital ad campaign performance by **predicting user click behavior** using **machine learning**. The goal is to **boost Click-Through Rate (CTR)** from 50% to 65% and **cut Cost Per Click (CPC)** by 20%.

Analyzing **1,000 user interactions**, **Logistic Regression** was chosen for its accuracy and interpretability, then fine-tuned for optimal results. Key insights show the **most responsive audience** is **aged 37–61**, with **low to middle income**, **short website visits**, and **light internet usage**. **West Java** and **Jakarta** stand out as high-potential regions.

The model achieved **98.33% accuracy**, **99.32% precision**, and **99.34% AUC**, with only **5 misclassifications** out of 300 test records. It's projected to raise **CTR to 99.32%** (+98.6%) and **reduce CPC by 49.6%**, far surpassing initial targets.

To validate these projections, a **1–2 week A/B testing** phase is recommended before full-scale implementation.

Source Code

LinkedIn

afsilmis@gmail.com