Ad Click Prediction: Data Engineering & EDA

Mohamed Haroun Mezned
University: Tek-Up
Group: ING-4-SDIA
Tutor: Haythem Ghazouani

Academic Year 2025-2026

Contents

1 Data Source & Context

1.1 Project Overview

The objective of this project is to predict whether a user will click on an online advertisement based on their demographic profile and browsing behavior. This phase focuses on transforming a static internal dataset into a dynamic, context-aware dataset suitable for Machine Learning.

1.2 Internal Dataset

The primary data source is `ad_click_dataset.csv`, containing the following key features:

- **Target Variable:** `click` (Binary: 0 = No Click, 1 = Click).
- **Demographics:** `age`, `gender`, `area_income`.
- **User Behavior:** `device_type`, `ad_position`, `browsing_history`, `time_of_day`.

1.3 Web Scraping (External Enrichment)

To capture the real-time context of user behavior, we implemented a web scraper targeting the *CNBC Technology* section.

1.3.1 Strategy

We extract trending headlines to understand what topics are currently popular (e.g., "AI", "Apple", "Stocks"). This allows us to determine if a user's browsing history aligns with current market trends.

```python
import requests
from bs4 import BeautifulSoup

def get_cnbc_trending_keywords():
    url = "https://www.cnbc.com/technology/"
    # Mimic browser headers to avoid blocking
    headers = {'User-Agent': 'Mozilla/5.0 ...'}

    response = requests.get(url, headers=headers)
    if response.status_code == 200:
        soup = BeautifulSoup(response.content, 'html.parser')
        # Extract headlines and split into keywords
        # ... extraction logic ...
        return list_of_keywords
```

Listing 1: Scraping Logic

2 Feature Engineering

Using the scraped data and logic rules, we created three new features to enhance model performance and data richness.

2.1 Feature 1: Market Relevance (`is_trending`)

We cross-referenced the user's `browsing_history` with the scraped CNBC keywords.

- **Logic:** If the user's history (e.g., "Shopping") matches a trending keyword (e.g., "Deals"), `is_trending = 1`.
- **Purpose:** Captures real-time market demand and relevance.

2.2 Feature 2: Tech-Savvy Segmentation

We derived a behavioral segment based on `Age` and `Device Type`.

```python
def get_tech_savvy_status(row):
    age = row['age'] if not pd.isna(row['age']) else 35
    device = row['device_type']

    # Young + Mobile/Tablet = High Tech Savvy
    if age < 30 and device in ['Mobile', 'Tablet']:
        return 'High'
    # Older + Desktop = Low Tech Savvy
    elif age > 50 and device == 'Desktop':
        return 'Low'
    else:
        return 'Medium'
```

Listing 2: Tech-Savvy Logic

2.3 Feature 3: Seasonality (`is_holiday_today`)

Recognizing that consumer behavior changes significantly during holidays, we utilized the `holidays` library.

- **Context:** Specific to Tunisia (TN).
- **Purpose:** Flagging if the log data was captured during a holiday period (e.g., Eid al-Fitr).

The result is an enriched dataset saved as `ad_data_enriched.csv`.

3 Exploratory Data Analysis (EDA)

With the enriched dataset, we now perform a comprehensive visual analysis to identify patterns, check data quality, and prepare for Phase 2 (Machine Learning).

3.1 Data Quality & Missing Values

Analysis: Figure 1 reveals missing data in `age`, `gender`, and `device_type`.
Implication for Week 2: We cannot simply drop these rows. We will utilize `SimpleImputer` within a Scikit-Learn Pipeline (using Median for numerical data and Most-Frequent for categorical data).

3.2 Target Distribution

Analysis: We check the balance of the `click` variable.
Implication for Week 2: If imbalance is detected (e.g., 90% No Clicks), we are required to use **SMOTE** (Synthetic Minority Over-sampling Technique) to balance the training set before modeling.

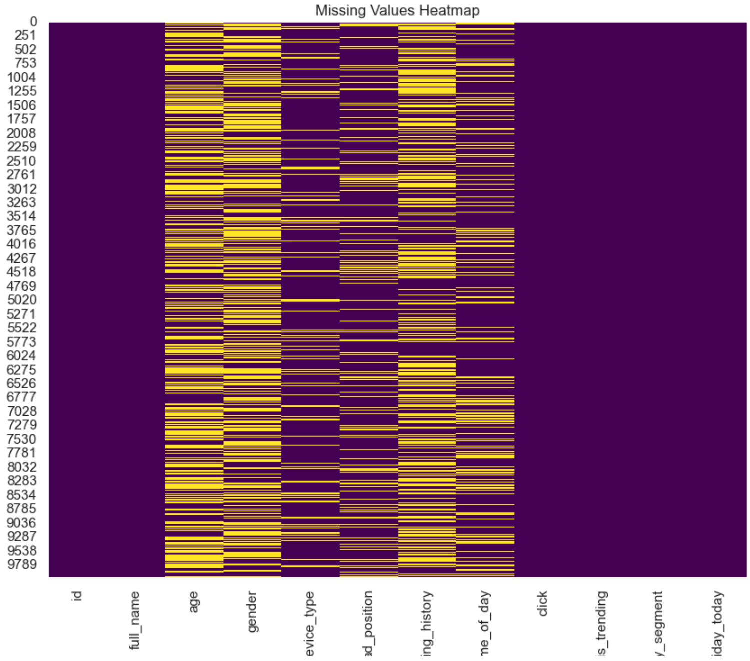

Figure 1: Missing Values Heatmap

3.3 Impact of New Features

3.3.1 Tech-Savvy Segmentation

Analysis: Comparing "High", "Medium", and "Low" segments allows us to see if younger, mobile-first users are more prone to clicking ads than older desktop users. If the "High" segment shows a higher Click-Through Rate (CTR), our feature engineering strategy is validated.

3.3.2 Trending Topic Influence

Analysis: Figure 4 validates the web scraping effort. If the "Trending" group has a significantly higher click rate than the "Not Trending" group, it confirms that external sentiment correlates with user action.

3.4 Correlation Analysis

Analysis: The heatmap helps identify strong predictors. Features highly correlated with `click` will likely have high feature importance in our XGBoost model in the next phase.

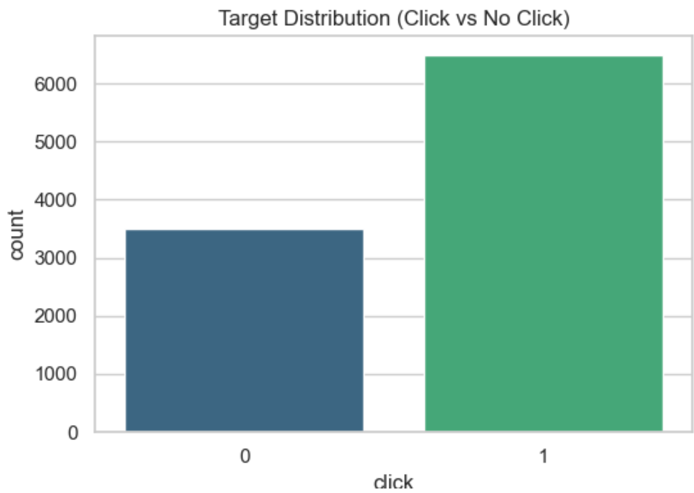

Figure 2: Target Distribution (Click vs No Click)

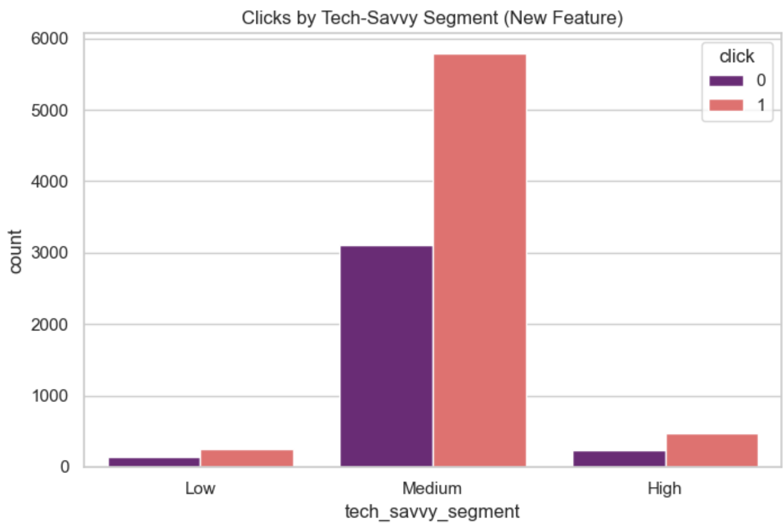

Figure 3: Click Rate by Tech-Savvy Segment

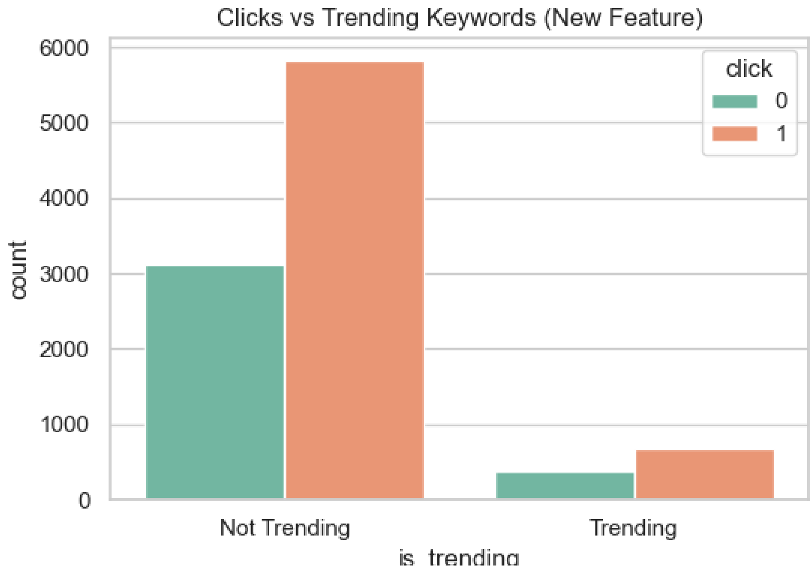

Figure 4: Clicks vs Trending Topics

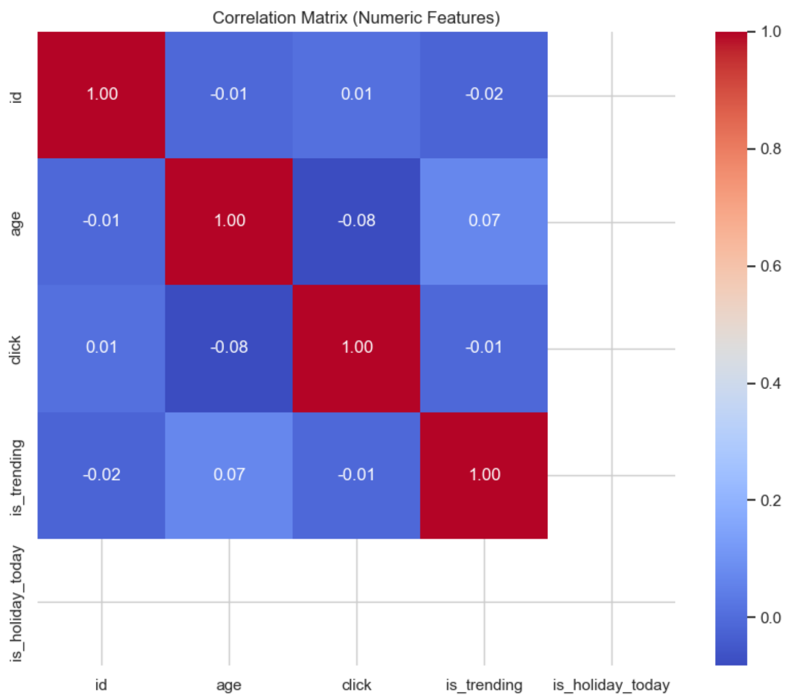

Figure 5: Correlation Matrix