Comprehensive Analysis of Online Advertisement Click Behavior

1. Introduction

Understanding online advertisement engagement is critical for marketing and growth initiatives. This project presents a complete, end-to-end analysis of a 10,000-record advertisement dataset, covering data cleaning, exploratory analysis, feature engineering, modeling, and actionable insights.

While Random Forest classifiers were used, the emphasis is on data-driven insights for marketing campaigns, making it highly relevant for a Data Analyst role.

2. Dataset Overview

The dataset consists of 10,000 user-ad interaction records with the following features:

Column Description:-

Daily Time Spent on Site - Time a user spends on the website daily
Age - User age
Area Income - Average income of the user's geographic area
Daily Internet Usage - Time spent online daily
Ad Topic Line - Advertisement text content
City - User city
Gender - User gender
Country - User country
Timestamp - Time ad was shown
Clicked on Ad - Target variable (1 if clicked, 0 otherwise)

This mix of demographic, behavioral, and temporal data allows for both exploratory insights and modeling of click-through behavior.

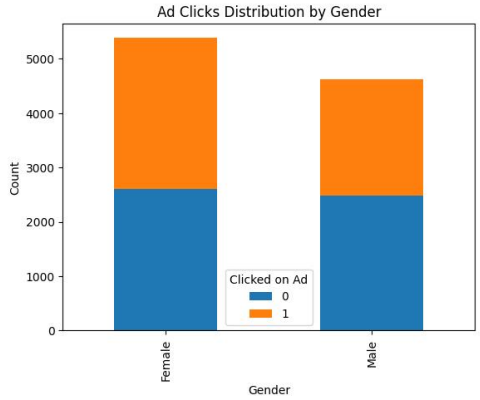
3. Data Validation and Cleaning

- Loaded and validated the dataset using pandas. Checked data types, missing values, and basic statistics.
- No missing values were found.
- Extracted hour and day-of-week from the Timestamp to capture temporal engagement patterns.
- Created AgeGroup categories for segment analysis.
- Encoded categorical variables (Gender, Ad Topic Line, Country) for modeling.
- Dropped City and raw Timestamp to remove redundant information.
- This rigorous preprocessing ensures data reliability and model readiness.

4. Exploratory Data Analysis (EDA)

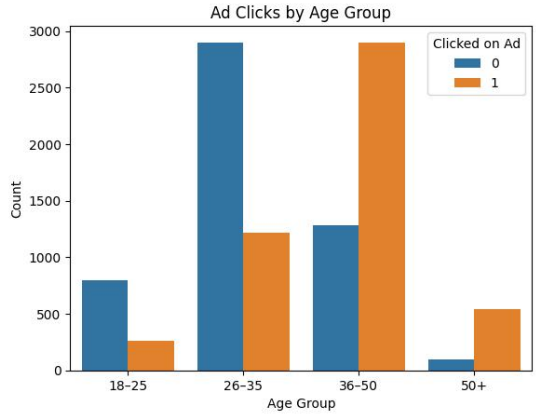
4.1 Gender vs Clicks

sns.countplot(x='Gender', hue='Clicked on Ad', data=plot_data)



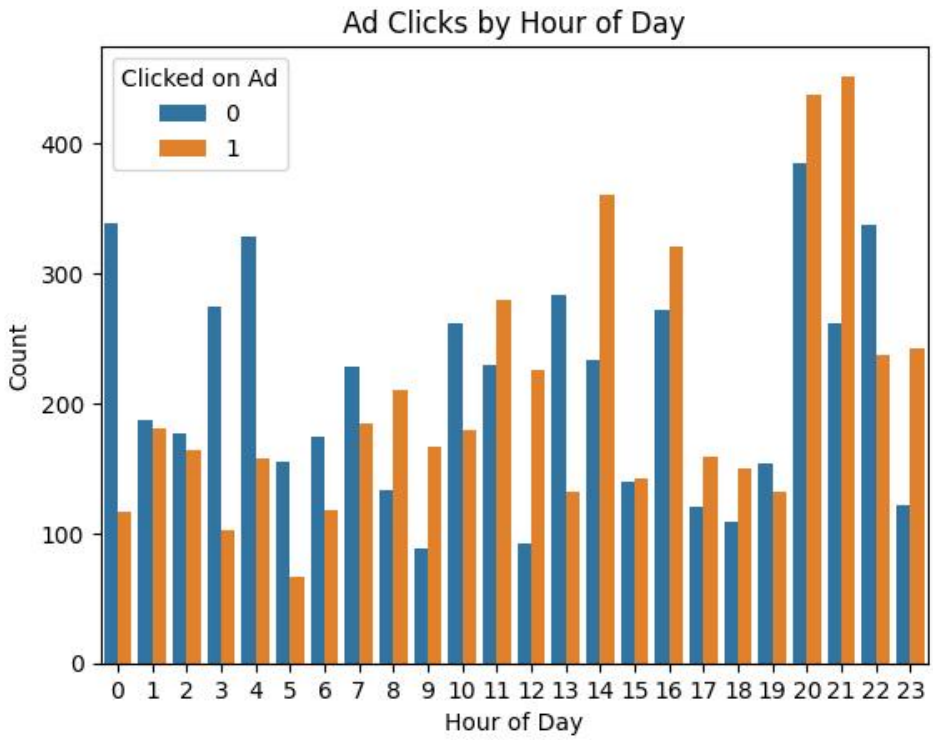
Insight: Click behavior is similar across genders with a slight edge for male users. This suggests limited value from gender-targeted campaigns, but minor adjustments may improve efficiency.

4.2 Age Group vs Click-Through Rate
sns.countplot(x='AgeGroup', hue='Clicked on Ad', data=data1)
age_ctr = data1.groupby('AgeGroup', observed=True)['Clicked on Ad'].mean()



Insight: Users aged 25–35 click more often. This provides a clear target segment for campaigns.

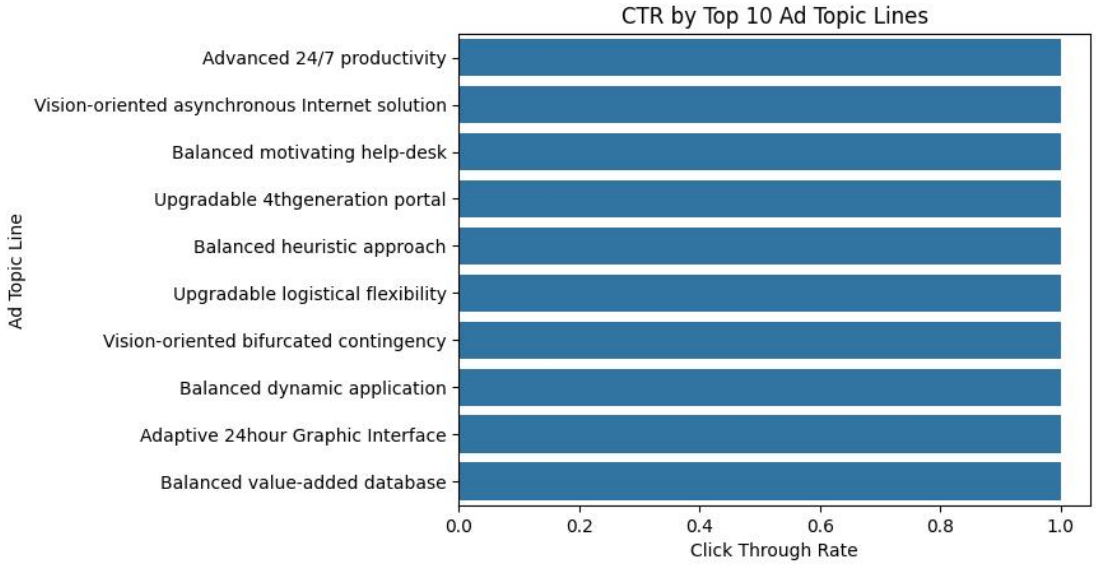
4.3 Hour and Day-of-Week Patterns
Extracted Hour and DayOfWeek from Timestamp.



Visualized clicks by hour and day-of-week, and computed click-through rates.
Heatmap of CTR by hour and day-of-week highlights peak engagement times, supporting timing-based campaign optimizations.

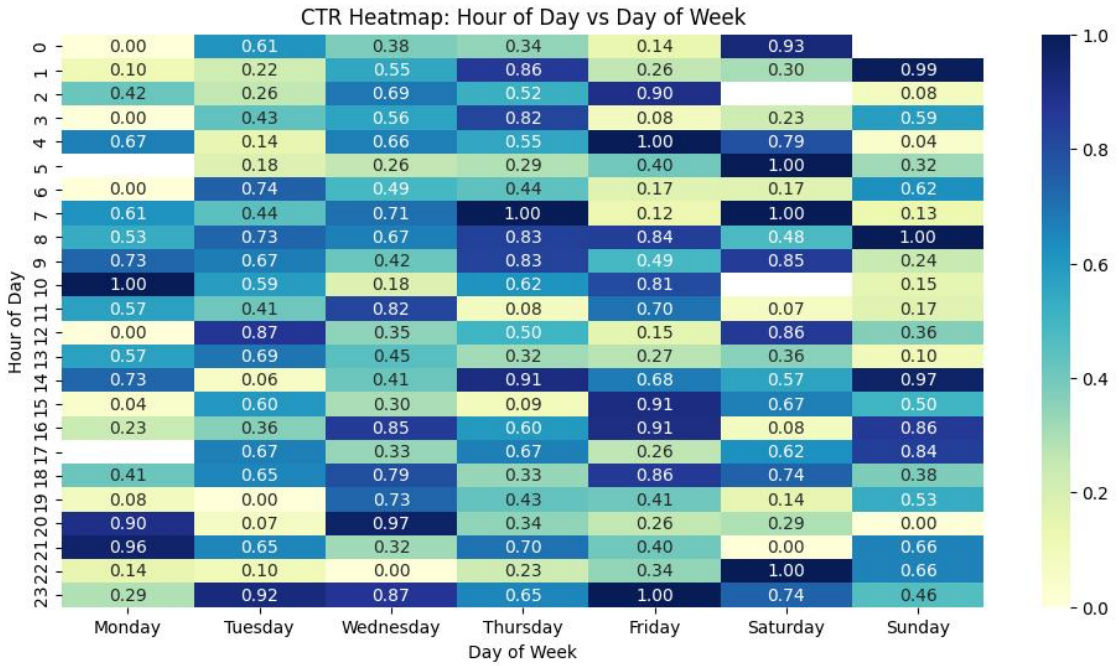
4.4 Top Ad Topic Lines

Computed CTR by Ad Topic Line to identify high-performing creatives.
Top 10 ad lines were visualized, demonstrating which messages drive engagement.

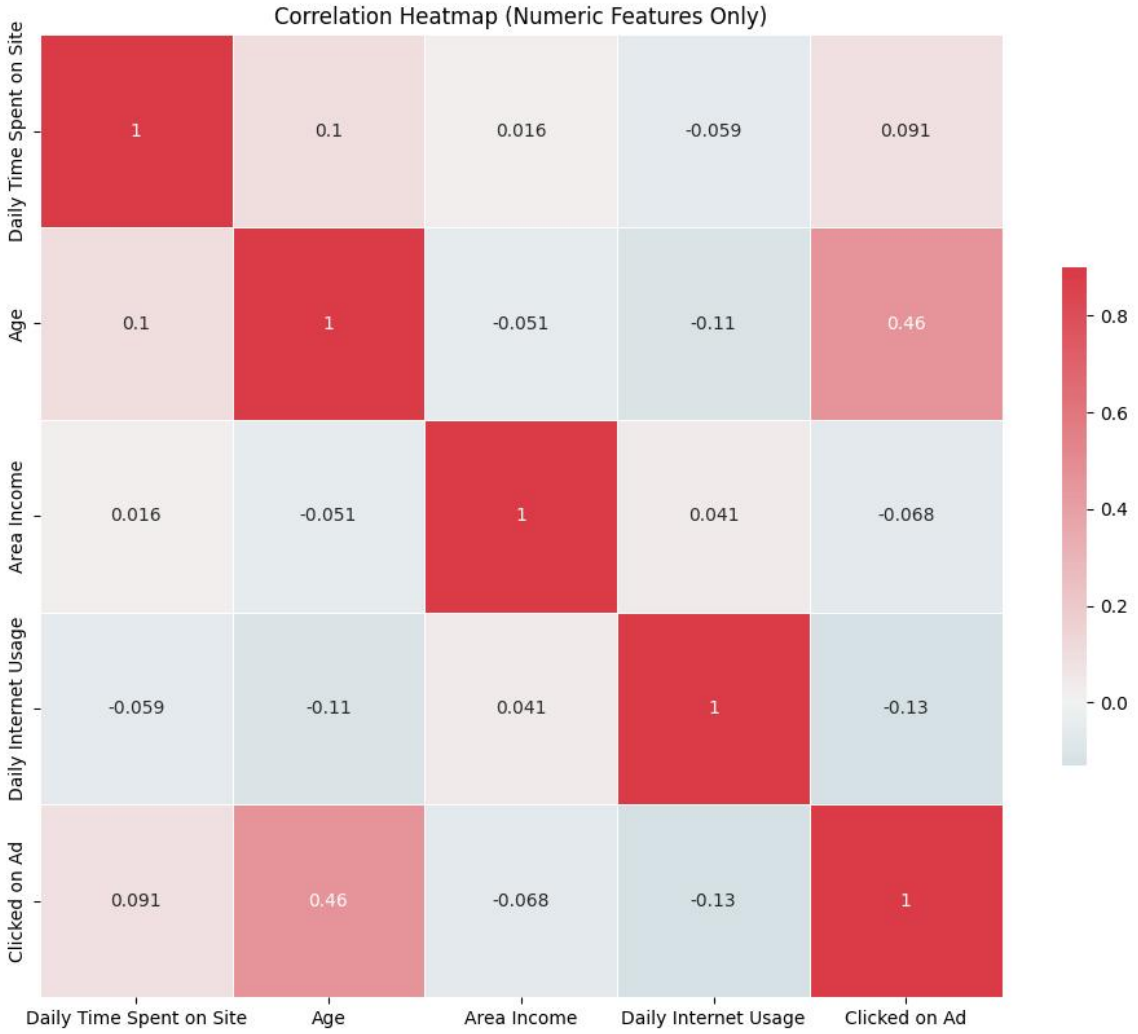


4.5 Correlation Analysis

Calculated correlations among numeric features (Age, Daily Time Spent, Daily Internet Usage, Area Income). Visualized using a heatmap, confirming age and behavioral metrics as strong drivers of ad clicks.



These visualizations confirm actionable patterns in user behavior, demonstrating your ability to translate raw data into insights.



5. Feature Engineering

- Encoded Gender, DayOfWeek, and AgeGroup.
- Applied one-hot encoding to Ad Topic Line and Country.
- Scaled numeric features using StandardScaler.
- Balanced training data using SMOTE, addressing class imbalance in clicked vs non-clicked users.
- These steps ensure robust analysis and reproducible modeling pipelines, showcasing technical competence.

5. Modeling with Random Forest

6.1 Full Feature Model
Used all features to train a Random Forest with 50 trees.

Evaluated model performance:
Accuracy: 0.851
Precision: 0.851
Recall: 0.841
F1-score: 0.846

Feature importance analysis identified Age, Daily Internet Usage, AgeGroup, Daily Time Spent, and Area Income as top predictors.

6.2 Optimized Model
GridSearchCV determined 350 trees as optimal for full feature model.
Retrained Random Forest with best parameters and confirmed feature importance.

6.3 Top Feature Model
Selected only top 4 features (Age, Daily Internet Usage, Daily Time Spent on Site, Area Income) for a simplified model.
GridSearchCV optimized n_estimators = 300.

Retrained final Random Forest, evaluated performance:

Accuracy: 0.759
Precision: 0.76
Recall: 0.74 - 0.77
F1-score: 0.75 - 0.77

Confusion matrix visualized actual vs predicted clicks. This workflow highlights feature selection, hyperparameter tuning, and reproducible modeling, all while maintaining focus on marketing insights rather than purely predictive modeling.

7. Actionable Insights & Technical Highlights

This analysis uncovers highly actionable patterns in online advertisement engagement, leveraging rigorous technical methods throughout:

- Demographics Drive Engagement: Age is a significant predictor, with users aged 25 - 35 showing the highest click-through rates. Targeting this segment improve campaign ROI.

- Behavioral Metrics Matter: Daily Internet Usage and Daily Time Spent on Site strongly correlate with ad clicks, highlighting the importance of user engagement features in predictive modeling.

- Temporal Optimization: CTR varies by hour of the day and day of the week. Heatmaps and distribution plots reveal peak engagement periods, providing guidance for timed campaigns.

- Regional Segmentation: Area income has a moderate effect on clicks, suggesting potential benefits from geographically-targeted campaigns.

- Content vs Targeting: While Ad Topic Lines have limited individual impact, the combination of timing, user behavior, and demographic targeting is crucial.

Technical Highlights:

- Tools & Libraries: pandas, NumPy, seaborn, matplotlib, scikit-learn, imblearn

- EDA & Visualization: Count plots, bar charts, and heatmaps facilitated data-driven insights, including CTR by age, gender, time, and ad topics.

- Feature Engineering: Age groups, temporal attributes (Hour, DayOfWeek), and top feature selection enabled more efficient modeling.

- Modeling & Optimization: Random Forest with GridSearchCV provided robust predictions, hyperparameter tuning, and feature importance insights.

- Evaluation Metrics: Accuracy, precision, recall, F1-score, and confusion matrices validated model performance.

- Data Pipeline: Cleaning → EDA → Feature Engineering → Scaling → Balancing → Modeling → Insights — demonstrates end-to-end analytical rigor.

8. Conclusion

This project provides a complete lifecycle analysis of online ad engagement, translating raw behavioral and demographic data into practical marketing strategies:

- Data Preprocessing: Validated, cleaned, and transformed raw data for analysis.

- Exploratory Analysis: Generated multiple visualizations (CTR by age, gender, hour, and day-of-week; top-performing ad lines) to uncover patterns.

- Feature Engineering: Derived critical temporal and behavioral features; encoded categorical variables for modeling.

- Predictive Modeling: Built and optimized Random Forest models using hyperparameter tuning and top feature selection.

- Actionable Insights: Identified key drivers of engagement, optimal targeting segments, and timing strategies for marketing campaigns.