CPSC-4830

GRANIFY DATA-DRIVEN MARKETING STRATEGIES ANALYSIS

Angeli De Los Reyes

Nay Zaw Lin (Neo)

Meyliani Sanjaya

Javier Merino

OBJECTIVE

Understand User Behavior

Analyze features such as FEATURE_1 to FEATURE_5 to capture session-level interactions and behavioral patterns

Assess Marketing Campaigns

Compare response outcomes (RESPONSE) between CONTROL and AD_ID groups to evaluate the impact of different advertisements

Measure Ad Performance:

Identify which advertisements are associated with higher conversion rates and user engagement

Derive Actionable Insights

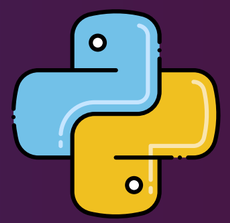
Use exploratory data analysis, data visualization, and predictive modeling to guide future marketing decisions

TOOLS FOR ANALYSIS

GitHub

Jupyter Notebook

Python



DATASET OVERVIEW

- **AD_ID**: Indicates which marketing strategy (Ad variant) was shown to the user (if any)
- **RESPONSE**: Target variable – 1 if the user responded (e.g., clicked or purchased), 0 otherwise
- **TIME**: Session timestamp in a 14-day marketing campaign, further broken down into DAY and HOUR for temporal analysis
- **CONTROL**: Binary indicator where 0 = user was shown an Ad (treatment group), 1 = user was not shown an Ad (control group)
- **FEATURE_1** to **FEATURE_5**: Categorical features representing various user/session 1 to 6 values attributes (e.g., user type, device, region, session length, products/images viewed, scroll speed, cart total, etc.)

DATA EXPLORATION

- Total observation: 14202 and Total Features: 9
- No missing values were found in the dataset across all columns
- Duplicate Records: 2,779 duplicate rows were detected

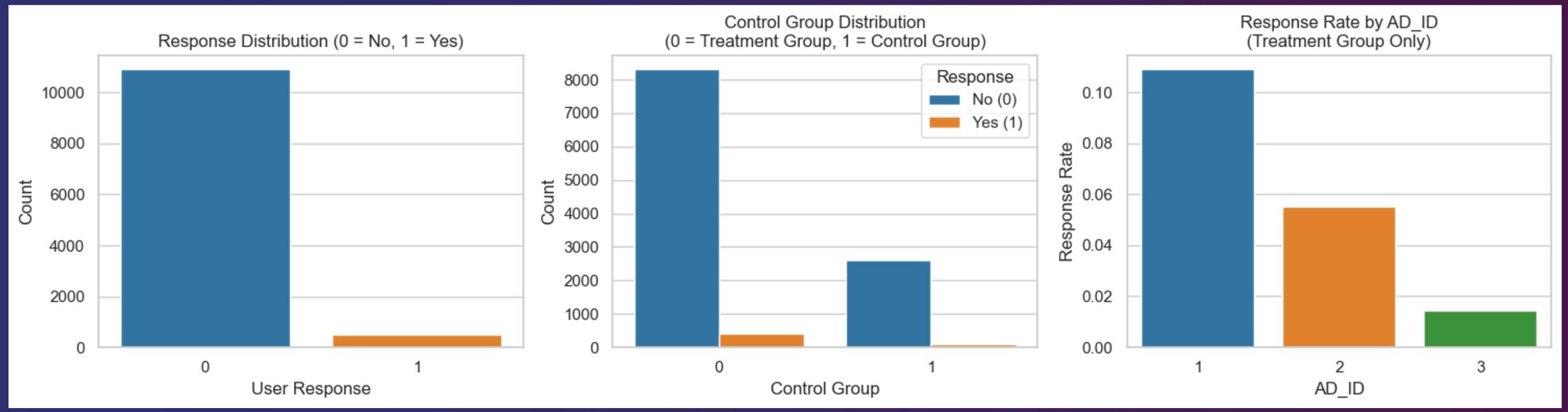
FEATURE DESCRIPTIONS TABLE

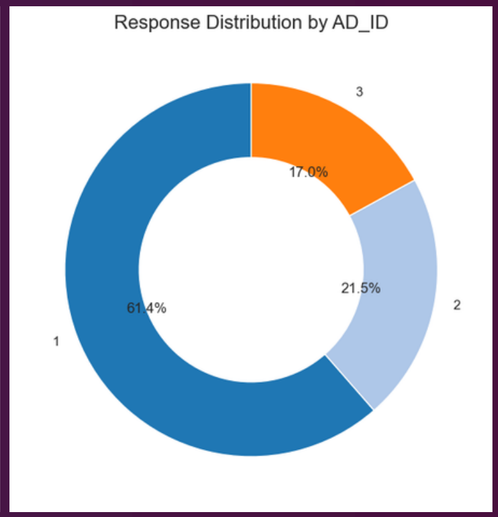
Feature	Data Type	Description
FEATURE_1	Categorical	Categorical variable describing a user/session attribute (e.g., device type)
FEATURE_2	Categorical	Categorical feature possibly related to demographics or user behavior
FEATURE_3	Categorical	Categorical session feature, likely from user interaction metadata
FEATURE_4	Categorical	Categorical variable with discrete levels (e.g., user interest group)
FEATURE_5	Categorical	Categorical session/user feature representing an attribute or category
AD_ID	Categorical	Identifier for which Ad strategy was shown (Range: 1, 2, 3)
CONTROL	Binary	Indicates whether user saw an Ad (`0` = Shown, `1` = Control group)
RESPONSE	Binary	Target variable, `1` if user responded positively, `0` otherwise
TIME	Categorical	Datetime String used to derive time-based insights such as DAY and HOUR. (Format: Day 1, 9:00)

DATA PREPROCESSING

- Remove the duplicate records
- Transform 'day' and 'hour' from the 'TIME' feature
- Features data type corrections

DATA VISUALIZATION
User Response and Ad Effectiveness Analysis

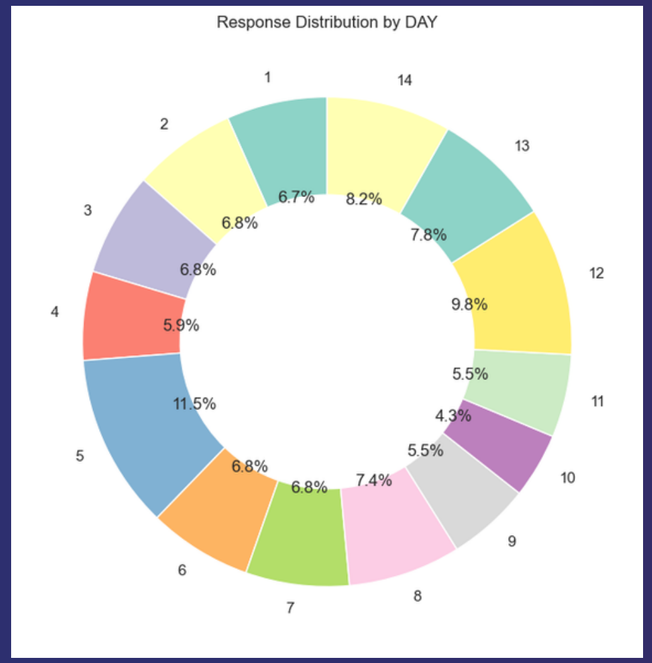


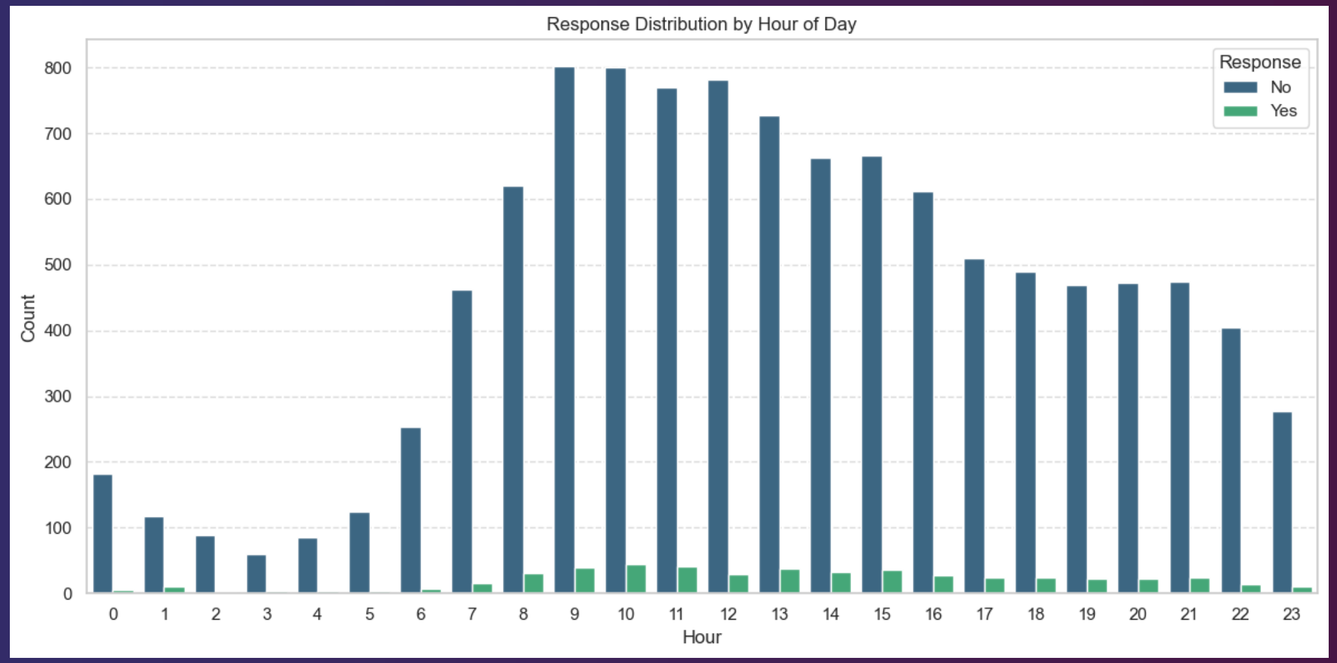


- **Class Imbalance**
  - Majority: RESPONSE = 0
  - Minority: RESPONSE = 1
  - Typical in marketing data
- **Low Conversion Rate**
  - Few users responded
  - Indicates weak persuasion or poor targeting
- **Control vs. Treatment**
  - Majority: Treatment (Ad shown)
  - Minority: Control (No ad shown)
  - Most users were shown ads (treatment group), while fewer users were in the control group
- **Ad Effectiveness (by AD_ID)**
  - Ad 1: Highest response (~11%), distribution (~61%) → Most persuasive
  - Ad 2: Moderate response (~6%), distribution (~21%)
  - Ad 3: Lowest response (~1%), distribution (~17%) → Needs revision
- **CONTROL 0 vs 1:**
  - CONTROL 1: This user was part of the control group – selected for a marketing strategy experiment but not shown one. Their behavior represents what would happen without the strategy's influence.
  - CONTROL 0: This user was in the treatment group – they were shown a specific ad (AD_ID 1, 2, or 3)

DATA VISUALIZATION

Time-Based Response Analysis





Day-Level Engagement

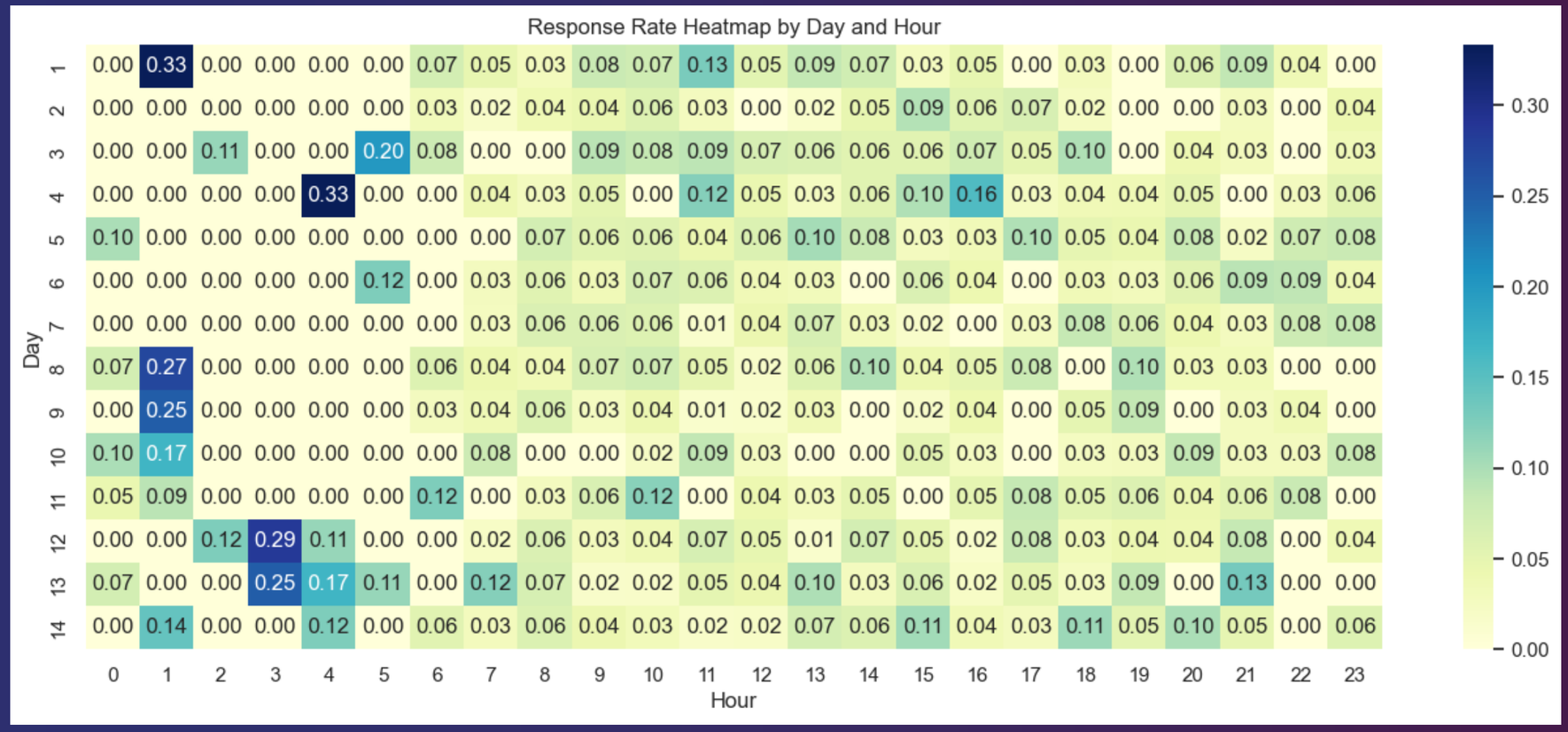
Higher responses on Day 5 and Day 12

Peak Hours

Most sessions between 9 AM – 3 PM

DATA VISUALIZATION

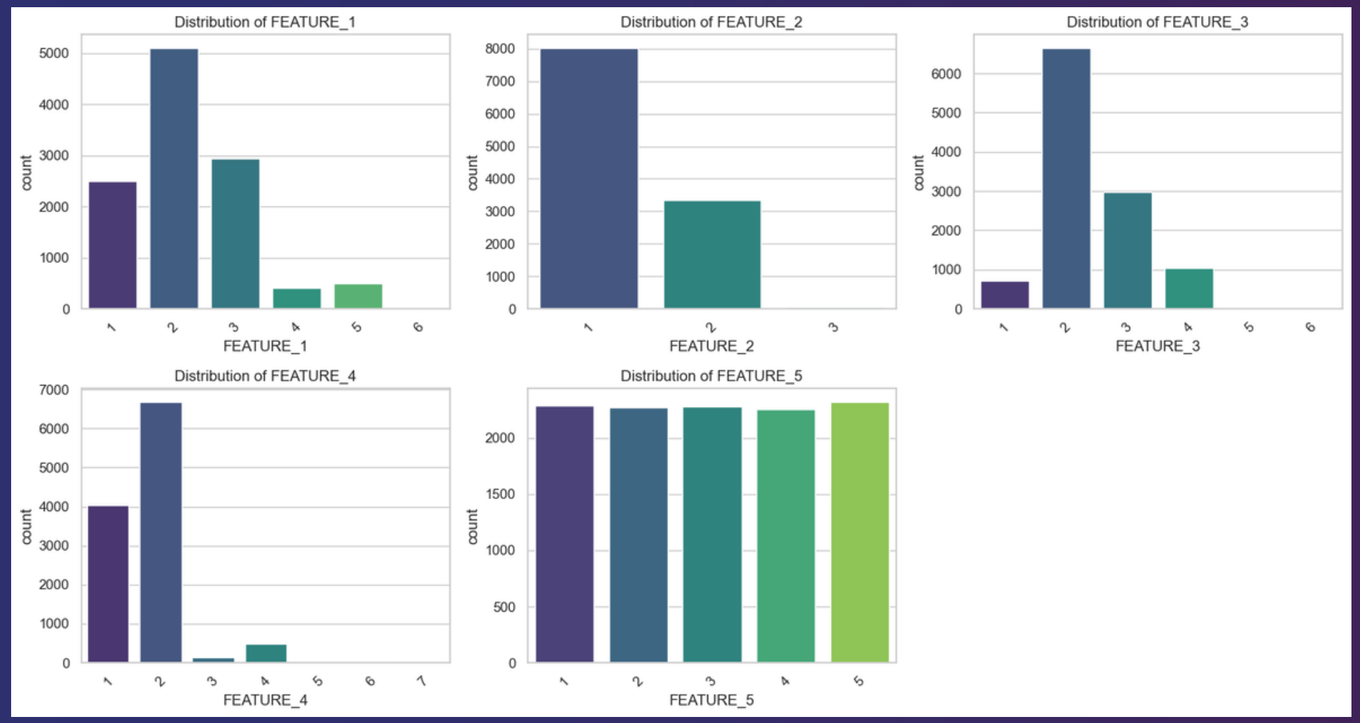
Time-Based Response Analysis

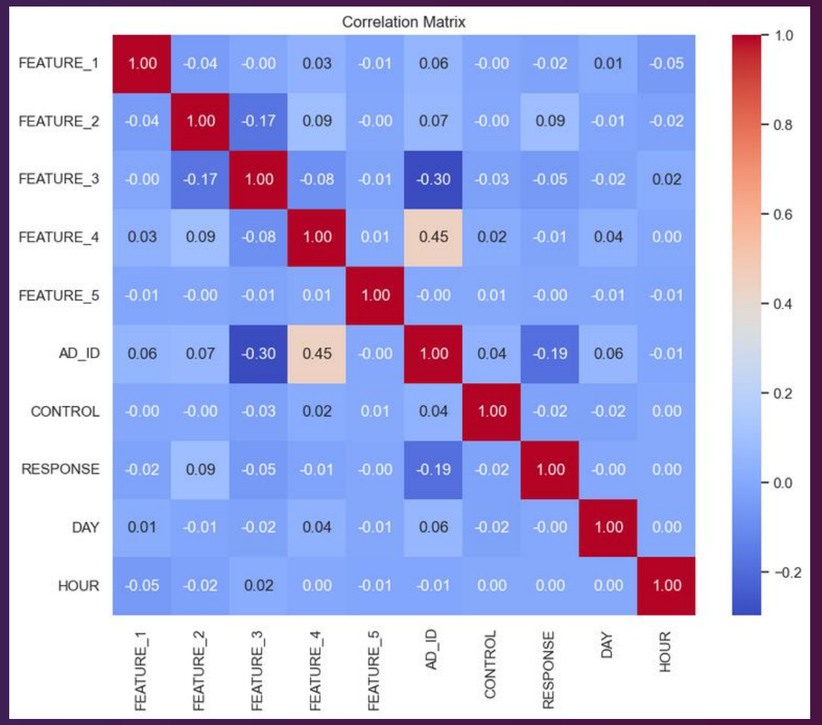


- Heatmap – Response Rate by Day and Hour
  - Day 1 at 1 AM and Day 4 at 4 AM – exceptionally high engagement (33%)
  - Day 8 at 1 AM, Day 9 at 1 AM, and Day 12 and 13 at 3 AM to 4 AM also show notable peaks (25–29%)

DATA VISUALIZATION

Feature Distributions and Correlation Analysis

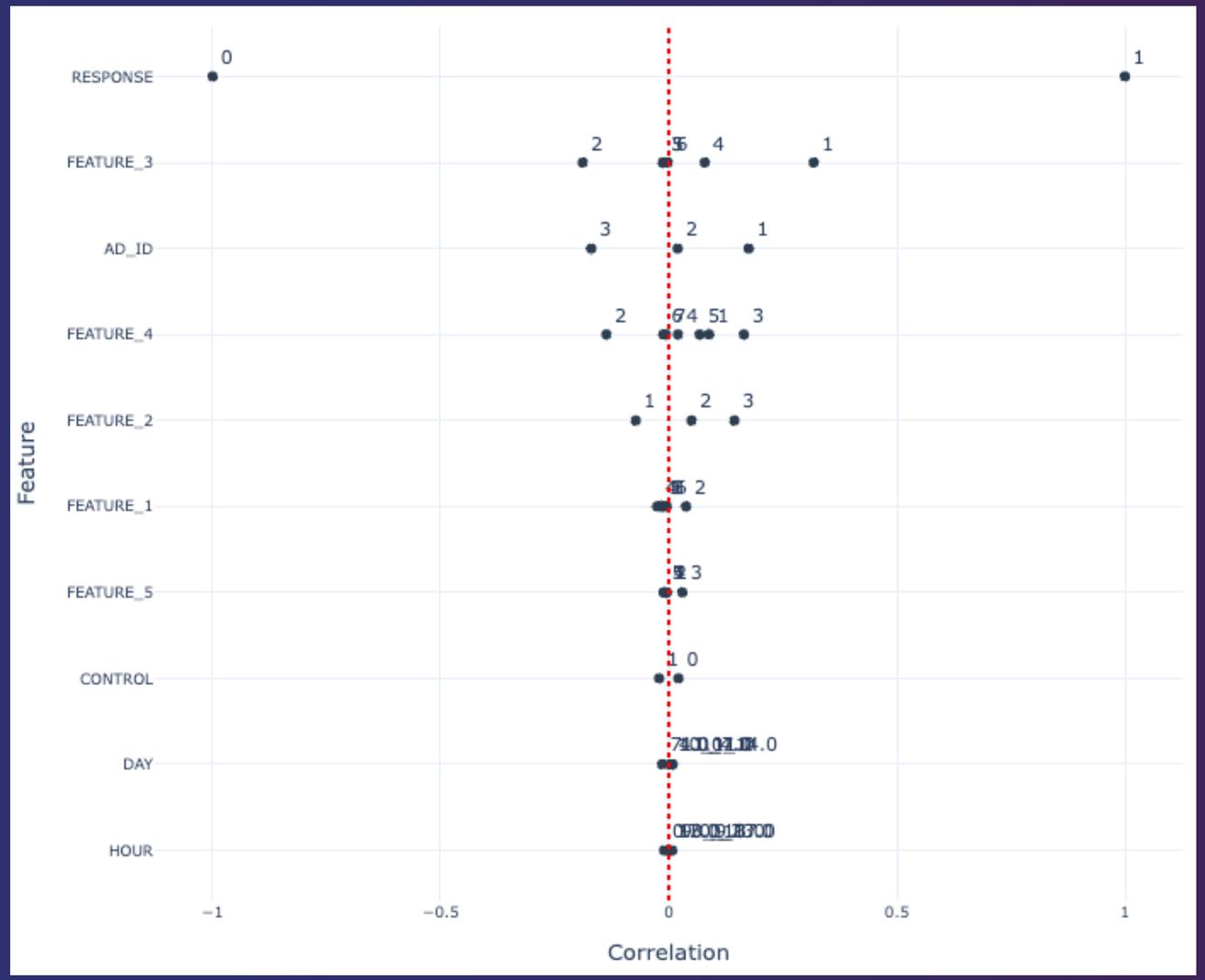




- Feature Distributions
  - FEATURE_1, FEATURE_3: Right-skewed, dominant categories
  - FEATURE_2: Heavily skewed to category 1
  - FEATURE_4: Imbalanced distribution
  - FEATURE_5: Uniform distribution → less predictive

- Correlation Insights
  - RESPONSE: Weakly correlated with all features
  - AD_ID:
    - Moderate correlation with FEATURE_4 (+0.45)
    - Negative correlation with FEATURE_3 (−0.30)
  - CONTROL: Uncorrelated → confirms random assignment
  - DAY & HOUR: Low direct correlation but potential non-linear effects
  - Multicollinearity: Low → features are mostly independent

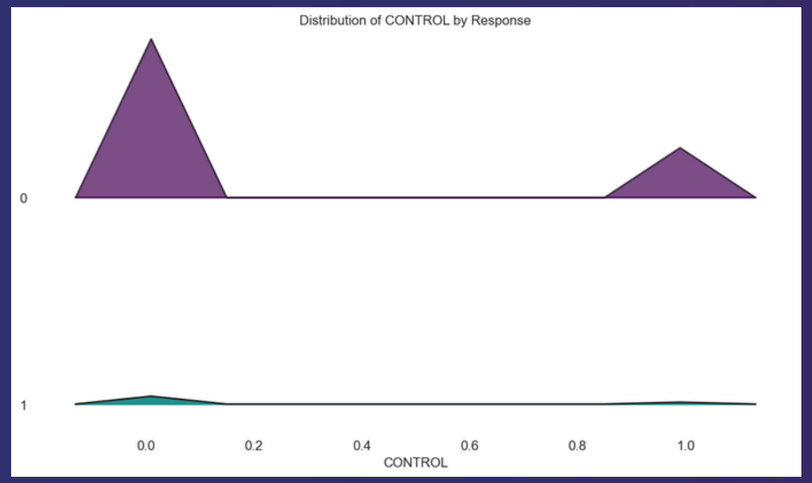
DATA VISUALIZATION
Correlation Funnel Visualization

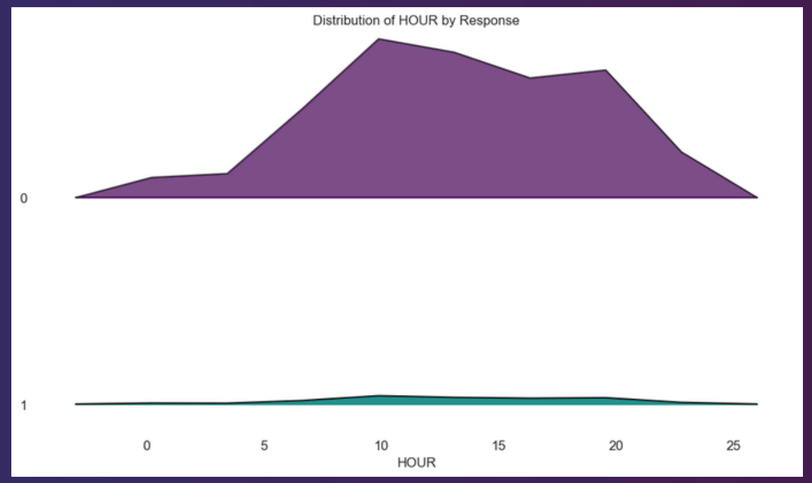


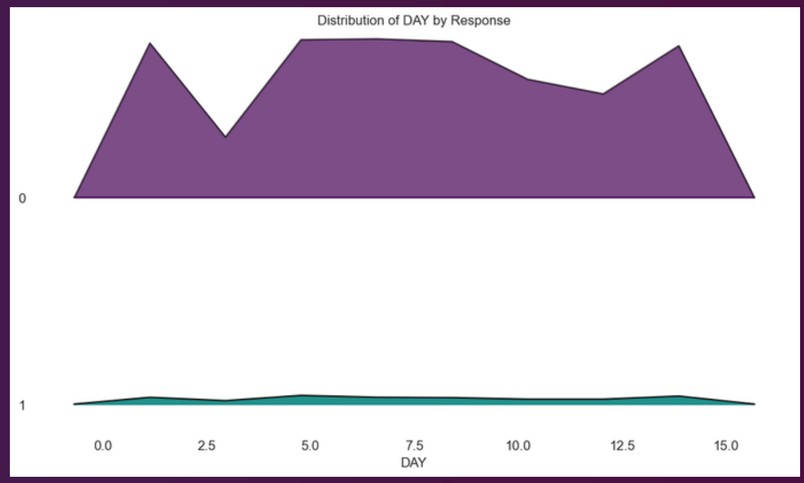
- **Goal**: Identify features that influence customer response to the marketing campaign (click/purchase).
- **Binary Correlation Analysis:** Convert continuous and categorical data to binary features and analyze their correlation with the target response.
- **Features with greater correlation with responding to the marketing campaign (right):**
  - For feature 3, category 1 is more likely to give a positive response.
  - For AD ID, ad strategy 1 is more likely to give a positive response.
  - For feature 4, category 3 is more likely to give a positive response.
  - For feature 2, category 3 is more likely to give a positive response.
  - The others have weak correlation.

DATA VISUALIZATION

Control, Time vs Response Visualization



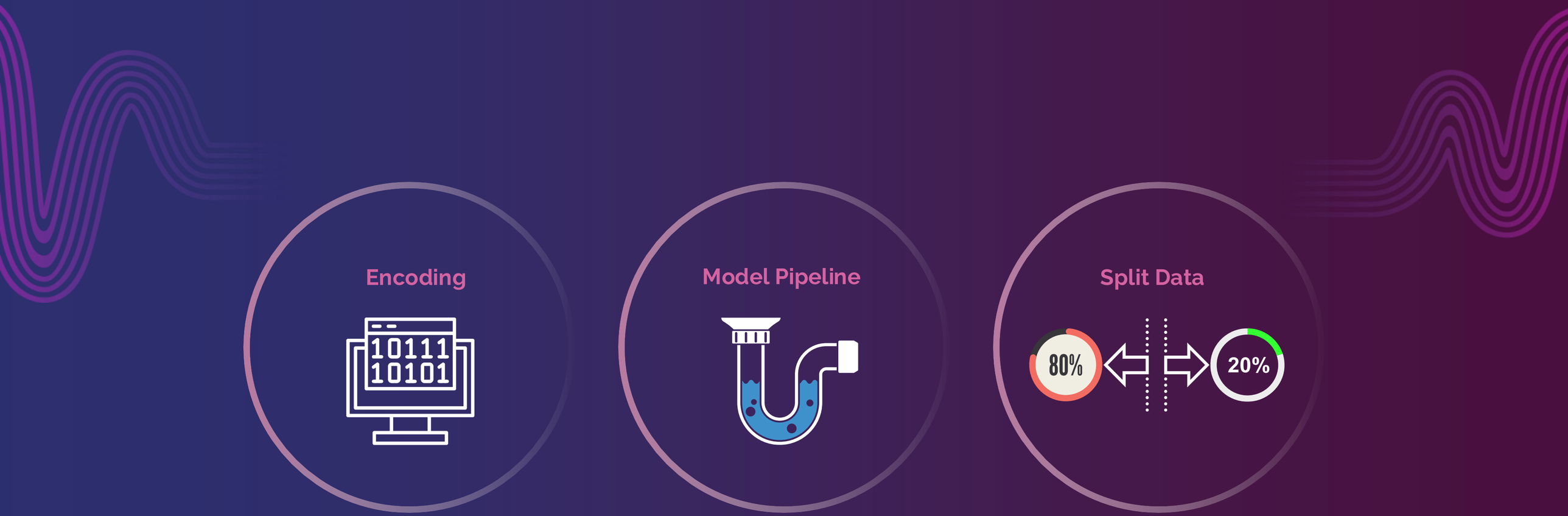




- Control Group
  - Nearly all responses came from the treatment group (CONTROL = 0)
  - Confirms Ad exposure drives user action
- Hour of Day
  - Non-responses peak between 10 AM–7 PM
  - Responses are spread out, slight bump at mid-day
  - Timing has some influence but not dominant
- Day of Campaign:
  - Responses are evenly distributed across days
  - DAY variable alone has low predictive power

Key Takeaway : who you show the ad to is matters

MODEL PREDICTIONS
Feature Engineering



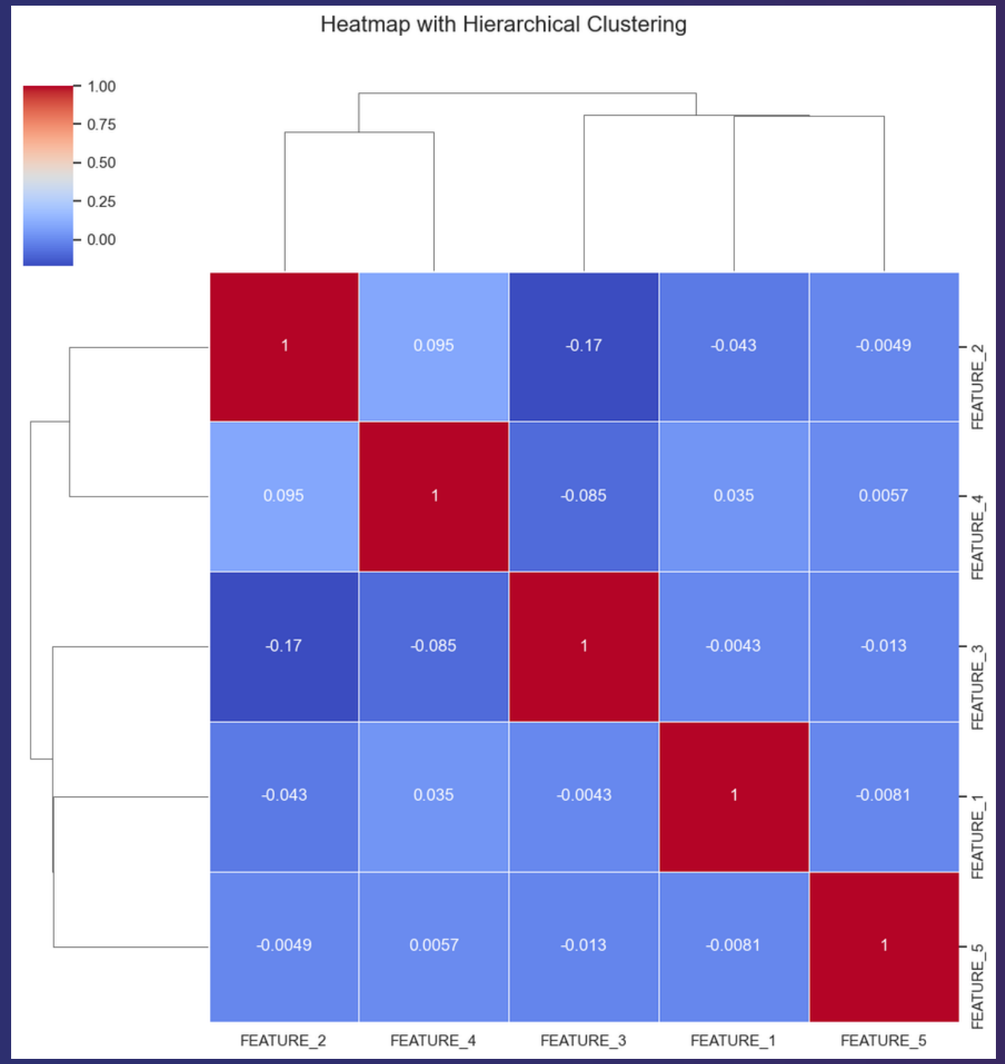
Encoding

Model Pipeline

Split Data

MODEL PREDICTIONS

Advanced Heatmap with Hierarchical Clustering



- Low Correlation:
  - Most features are weakly correlated, indicating they provide independent signals
- No Strong Clusters:
  - Dendrogram shows features are not grouped into tight clusters — each adds unique value
- Mild Associations Only:
  - Slight links exist (e.g., FEATURE_2 & FEATURE_4: +0.095; FEATURE_2 & FEATURE_3: -0.17), but not strong enough to indicate redundancy
- Modeling Implication:
  - Independence supports models like Random Forest or XGBoost
  - The dimensionality reduction approach is not critical

MODEL PREDICTIONS

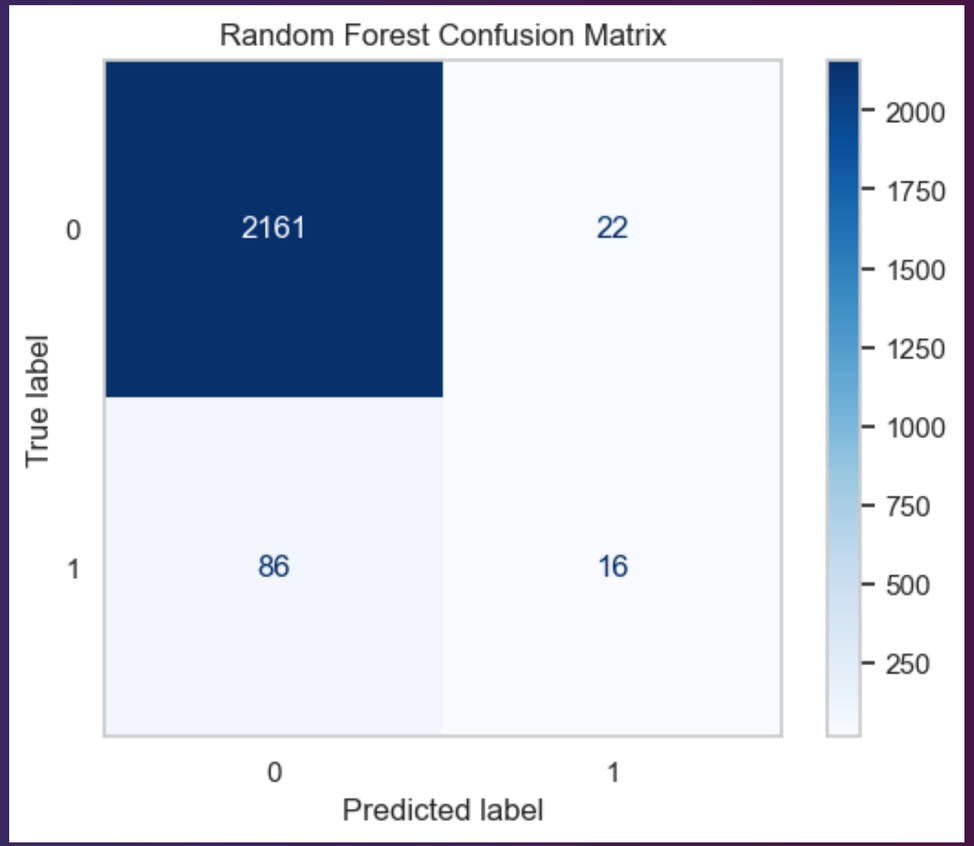
Model Building and Evaluation: Advanced Analysis (Random Forest)

```
Random Forest Accuracy: 0.95

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.98      2183
           1       0.42      0.16      0.23       102

    accuracy                           0.95      2285
   macro avg       0.69      0.57      0.60      2285
weighted avg       0.94      0.95      0.94      2285
```



- Overall Accuracy
  - 95% - strong overall classification
- Weighted average score:
  - Precision: 94%
  - Recall: 95%
- Confusion Matrix
  - False Negatives (86): Most responses were missed
  - True Positives (16): Few positive cases were identified correctly

MODEL PREDICTIONS

Logistic Regression, Random Forest, and XGBoost Comparisons

Logistic Regression Accuracy: 0.96

Random Forest Accuracy: 0.95

XGBoost Accuracy: 0.95

Logistic Regression Classification Report:

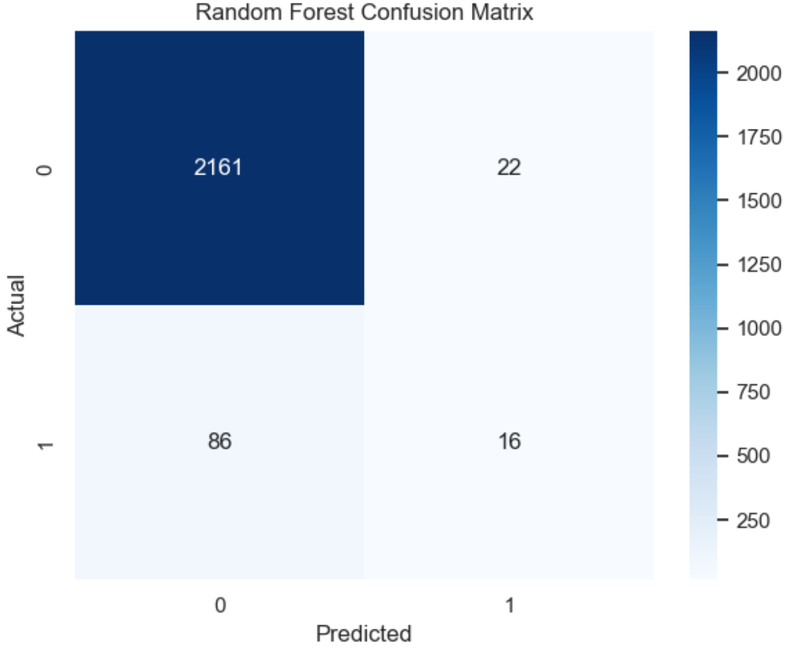
	precision	recall	f1-score	support
0	0.96	1.00	0.98	2183
1	0.50	0.01	0.02	102
accuracy			0.96	2285
macro avg	0.73	0.50	0.50	2285
weighted avg	0.94	0.96	0.93	2285

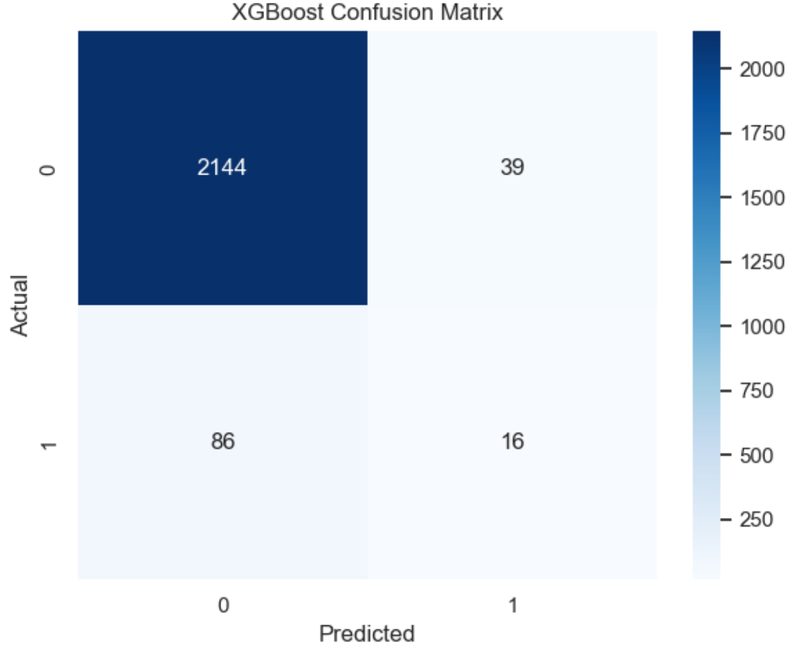
Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.96	0.99	0.98	2183
1	0.42	0.16	0.23	102
accuracy			0.95	2285
macro avg	0.69	0.57	0.60	2285
weighted avg	0.94	0.95	0.94	2285

XGBoost Classification Report:

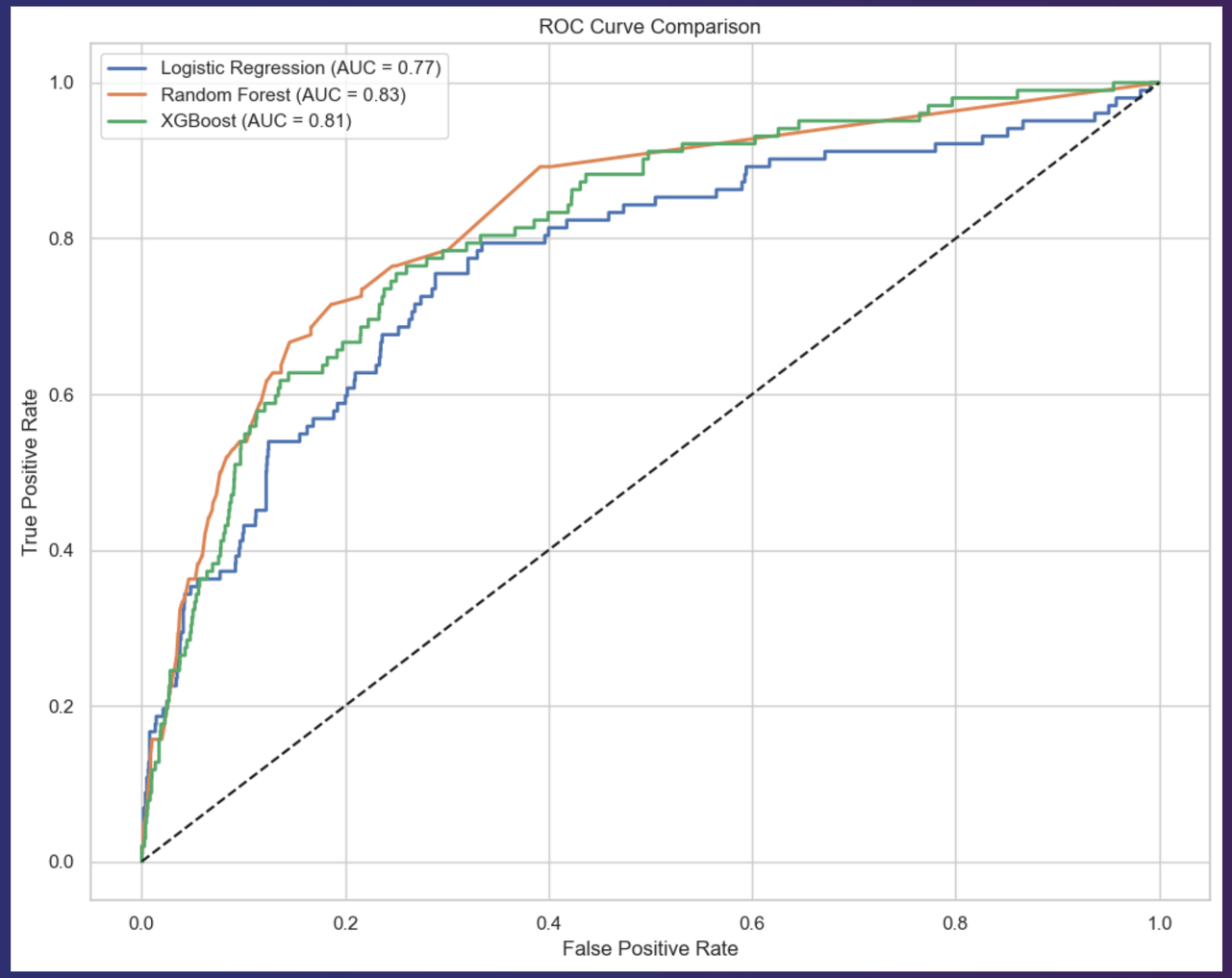
	precision	recall	f1-score	support
0	0.96	0.98	0.97	2183
1	0.29	0.16	0.20	102
accuracy			0.95	2285
macro avg	0.63	0.57	0.59	2285
weighted avg	0.93	0.95	0.94	2285





MODEL PREDICTIONS

Logistic Regression, Random Forest, and XGBoost Comparisons



- AUC Scores
  - Logistic Regression: 0.77
  - Random Forest: 0.83
  - XGBoost: 0.81
- Random Forest and XGBoost Outperform Logistic Regression in AUC Scores

MODEL PREDICTIIONS

SMOTE + Cross-Validation, Hyperparameter Tuning, Evaluation Analysis

Logistic Regression Accuracy: 0.71
Random Forest Accuracy: 0.94
XGBoost Accuracy: 0.94

Logistic Regression Classification Report:

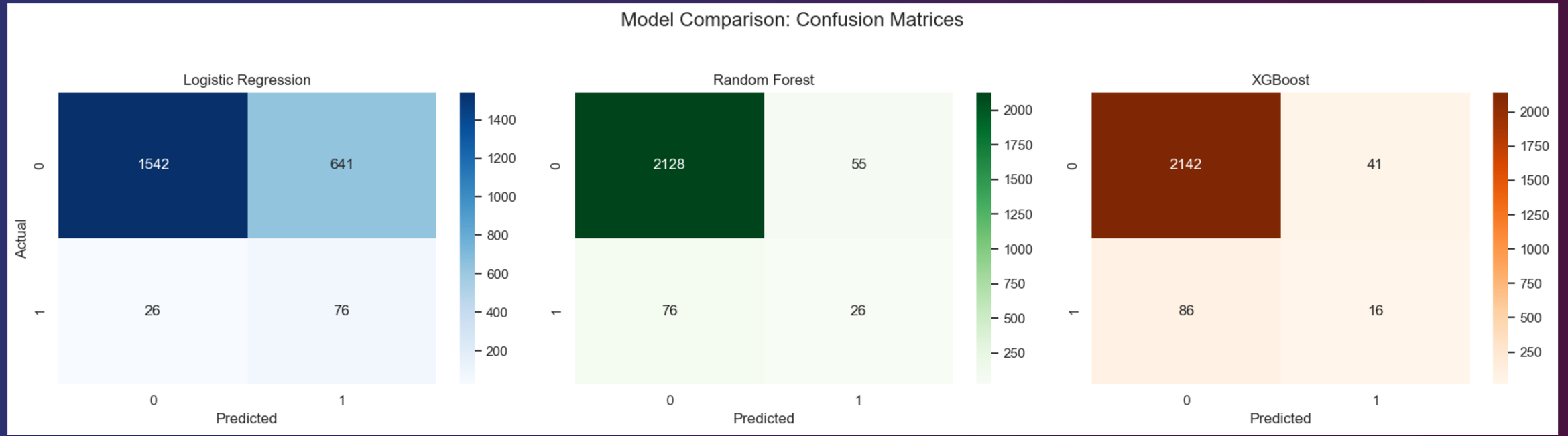
	precision	recall	f1-score	support
0	0.98	0.71	0.82	2183
1	0.11	0.75	0.19	102
accuracy			0.71	2285
macro avg	0.54	0.73	0.50	2285
weighted avg	0.94	0.71	0.79	2285

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.97	0.97	0.97	2183
1	0.32	0.25	0.28	102
accuracy			0.94	2285
macro avg	0.64	0.61	0.63	2285
weighted avg	0.94	0.94	0.94	2285

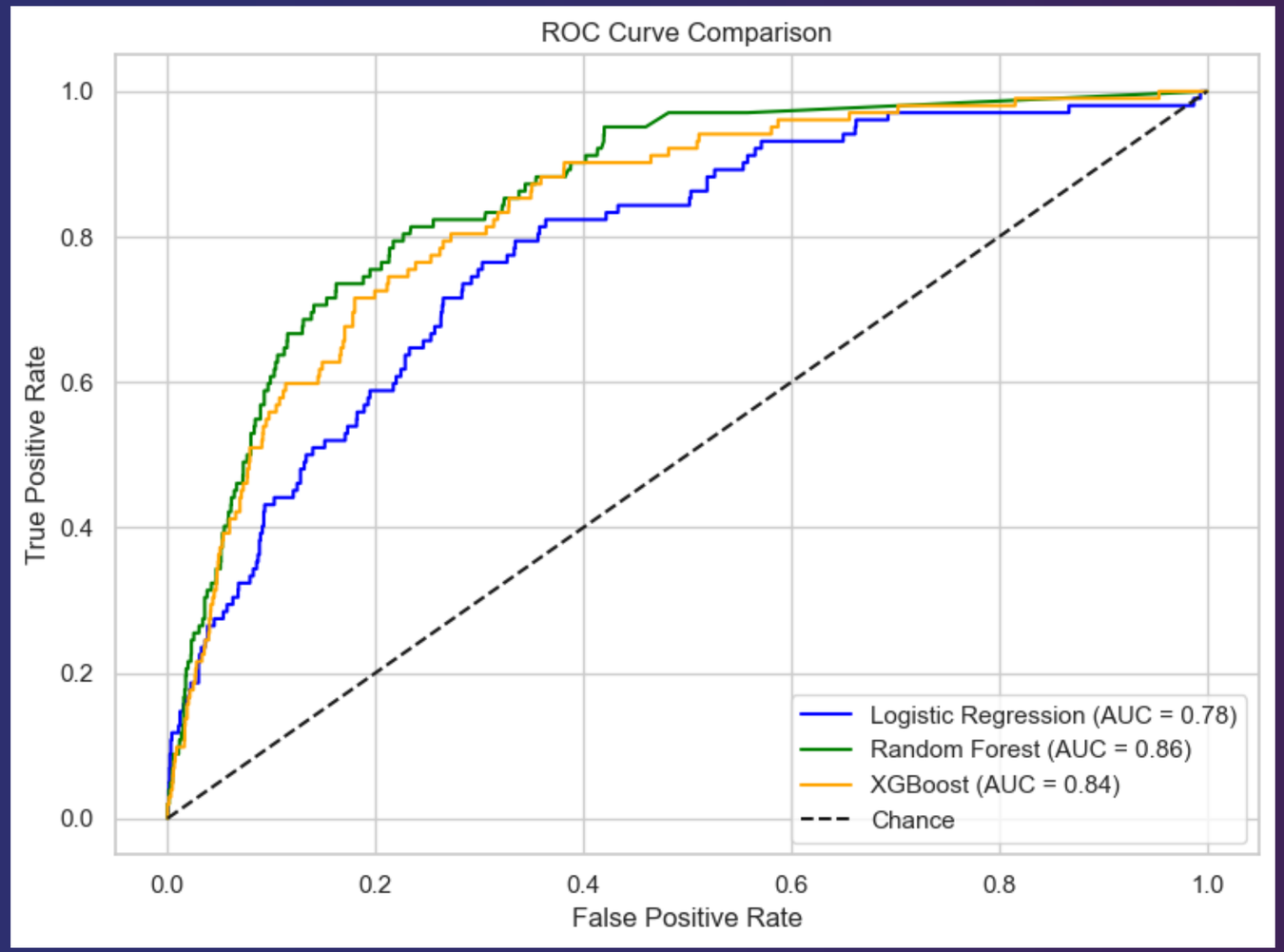
XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	2183
1	0.28	0.16	0.20	102
accuracy			0.94	2285
macro avg	0.62	0.57	0.59	2285
weighted avg	0.93	0.94	0.94	2285



MODEL PREDICTIIONS

Cross-Validation, SMOTE + Hyperparameter Tuning, Evaluation Analysis



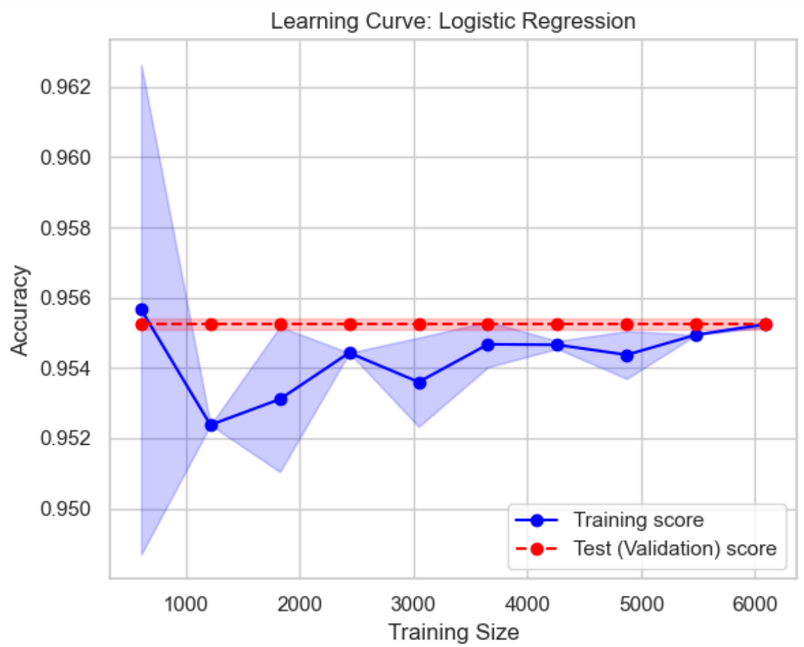
AUC Scores

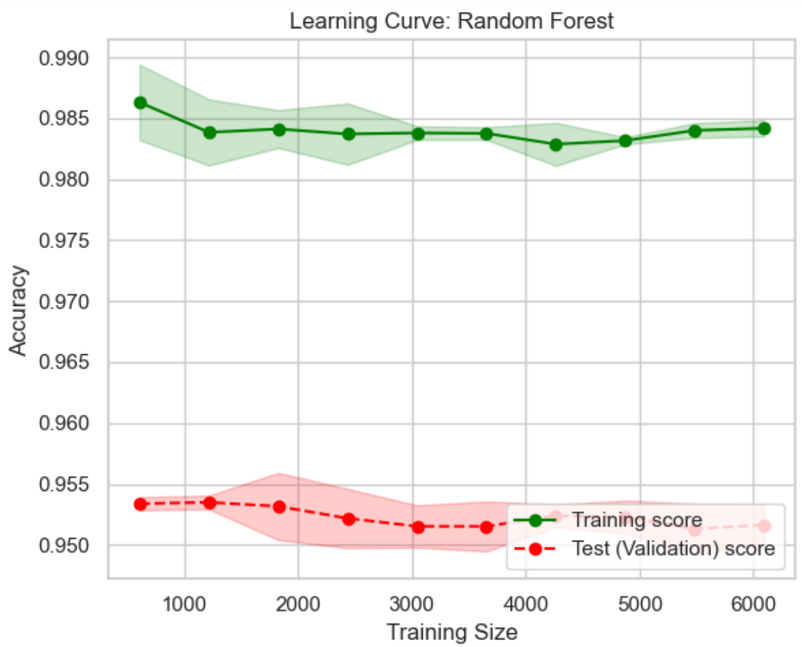
- Logistic Regression: 0.78
- Random Forest: 0.86
- XGBoost: 0.84

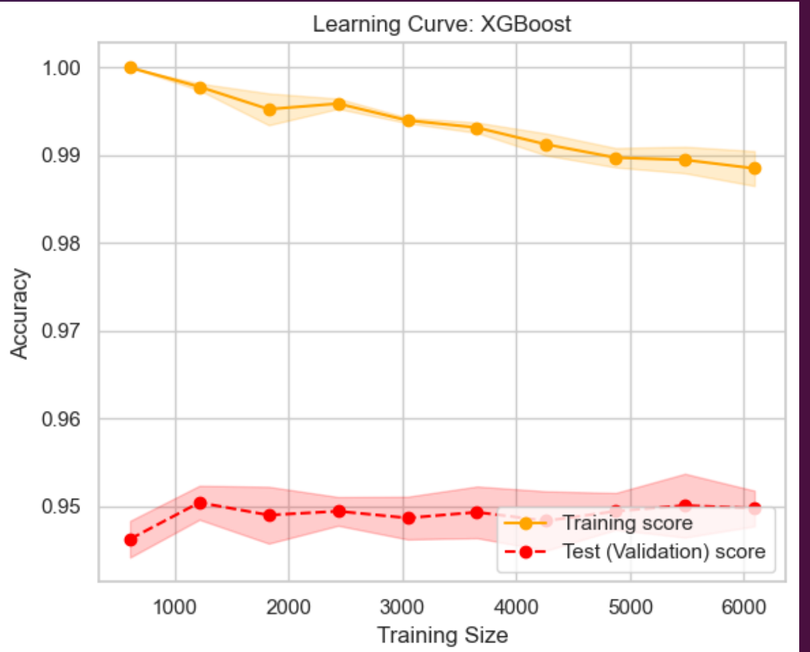
Random Forest and XGBoost Outperform Logistic Regression in AUC Scores

MODEL PREDICTIONS

Learning Curve for Logistic Regression, Random Forest and XGBoost Models







Model	Gap	Interpretation
Logistic Regression	Small	Stable and well-generalized.
Random Forest	Moderate	Slightly overfitting observed.
XGBoost	Large	Strong overfitting. Excellent fit on train, not on test.

MODEL PREDICTIONS

All models comparison

Model	Setting	Accuracy	Recall (Class 1)	AUC	Key Insight
Logistic Regression	Baseline	0.96	0.01	0.77	High accuracy, misses positives
Logistic Regression	SMOTE + Tuning	0.71	0.75	0.78	Best recall, good for targeting responders
Random Forest	Baseline	0.95	0.16	0.83	Balanced model, limited recall
Random Forest	SMOTE + Tuning	0.94	0.25	0.86	Strong AUC, improved minority detection
XGBoost	Baseline	0.95	0.16	0.81	Stable model, recall needs improvement
XGBoost	SMOTE + Tuning	0.94	0.16	0.84	Slight recall boost, consistent performance

- Random Forest with SMOTE and Hyperparameter Tuning
- Balancing accuracy (94%), recall (25%), and AUC (86%)

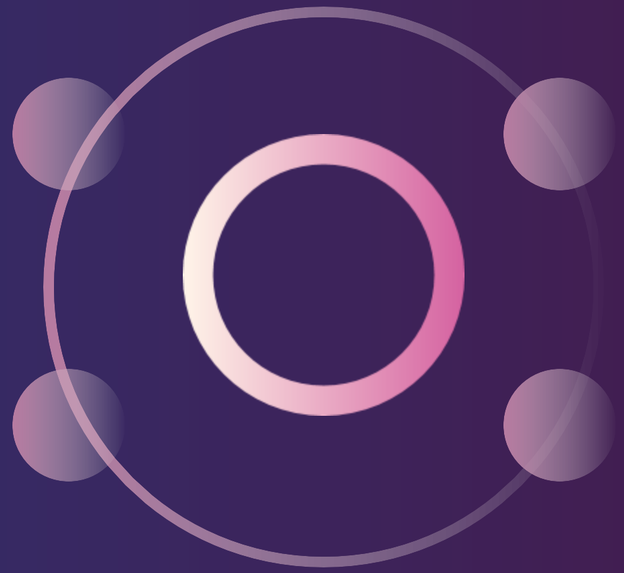
ANALYSIS CONCLUSION

FEATURES CORRELATION

The features in this dataset show low to moderate correlation, suggesting that each one provides relatively distinct behavioral or contextual information.

IMPORTANT FEATURES

Features like AD_ID, time (HOUR, DAY), and user/session traits (FEATURE_2, FEATURE_4) are crucial to model response likelihood and evaluate marketing strategies.



EFFECTIVE MARKETING STRATEGIES

Time-targeted marketing is effective - ads delivered in off-peak hours can outperform daytime when targeting the right segments result on increase conversion

User Segmentation: Identify and prioritize high-response user groups (e.g., by AD_ID or session patterns) for more personalized and effective marketing.

ACTIONABLE INSIGHT

- Run targeted ads during late-night windows (1–4 AM)
- Focus budgets on top-performing ads (e.g., AD_ID 1) and high-response periods to maximize return on investment (ROI).
- Utilize control groups to test new ad strategies.
- Audit or optimize underperforming AD_IDs
- Segment users based on FEATURE_x combos. Some combinations correlate with higher responsiveness

THANK YOU!