# “CLICKED ON AD”

USING PYTHON & MACHINE LEARNING

LIWU
DEC 21,2024

# AGENDA

1. Introduction & Business Problem
2. Data Description
3. Exploratory Data Analysis (EDA)
4. Feature Engineering
5. Model Development & Comparison
6. Business Recommendations

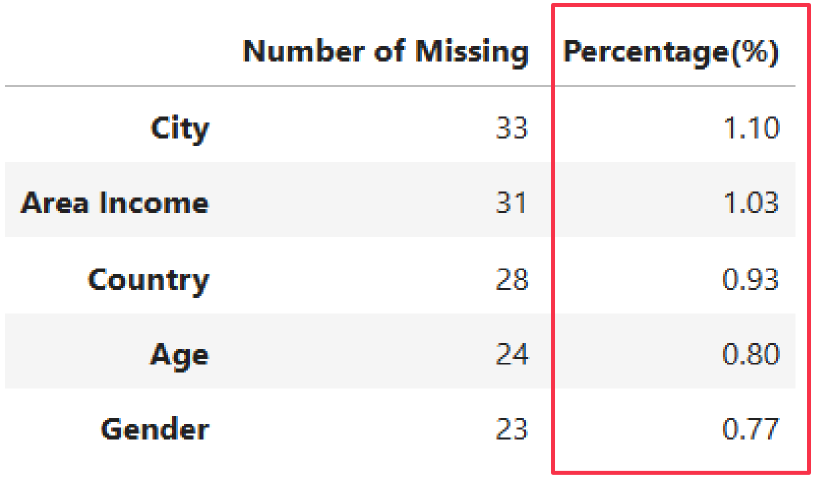
# INTRODUCTION & BUSINESS PROBLEM

**"How can we predict if a user will click on an advertisement?"**

- **Boosts CTR (Click-Through Rate)**: Encourages more user interactions with ads.
- **Enhances Targeting**: Focuses on likely-to-engage users.
- **Actionable Insights**: Refines marketing strategies effectively.
- **Increases ROI (Return on Investment)** : Maximizes profit from ad spend.
- **Better User Experience**: Shows relevant ads to users.

# DATA DESCRIPTION

- 2016 **Jan-July**
- **3,000** observations
- Duplicates: **31**
- Missing Values:

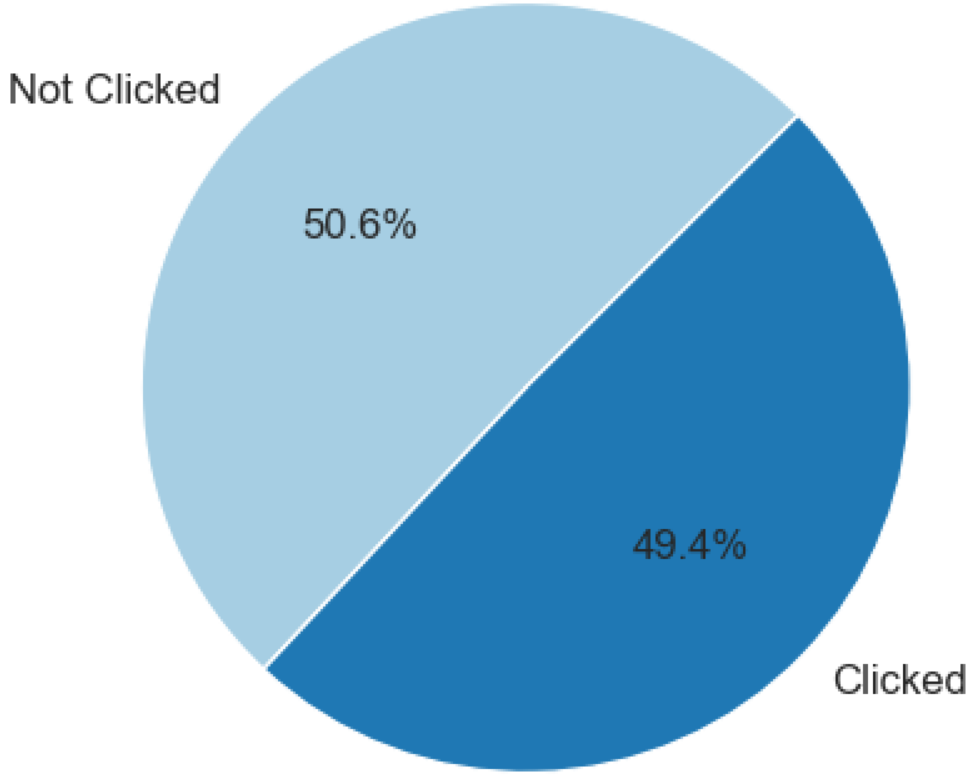


| | Number of Missing | Percentage(%) |
|---|---|---|
| **City** | 33 | 1.10 |
| **Area Income** | 31 | 1.03 |
| **Country** | 28 | 0.93 |
| **Age** | 24 | 0.80 |
| **Gender** | 23 | 0.77 |

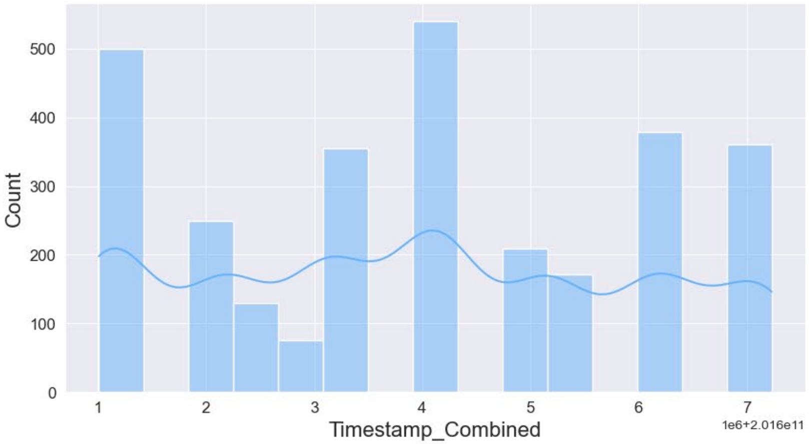
| **Target Variable** | Clicked on Ad |
|---|---|
| **Numerical Variables** | • Timestamp (Datetime)<br>• Daily Time Spend on Site<br>• Daily Internet Usage<br>• Age<br>• Area Income |
| **Categorical Variables** | • Gender<br>• Ad Topic Line<br>• City<br>• Country |

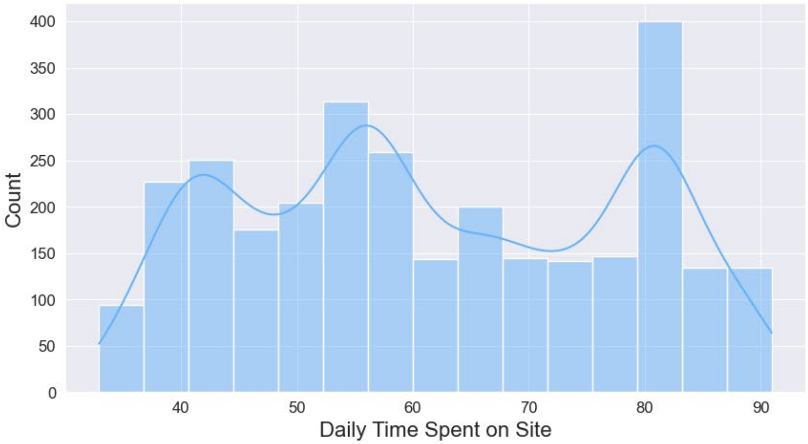
# EXPLORATORY DATA ANALYSIS

# “CLICKED ON AD”



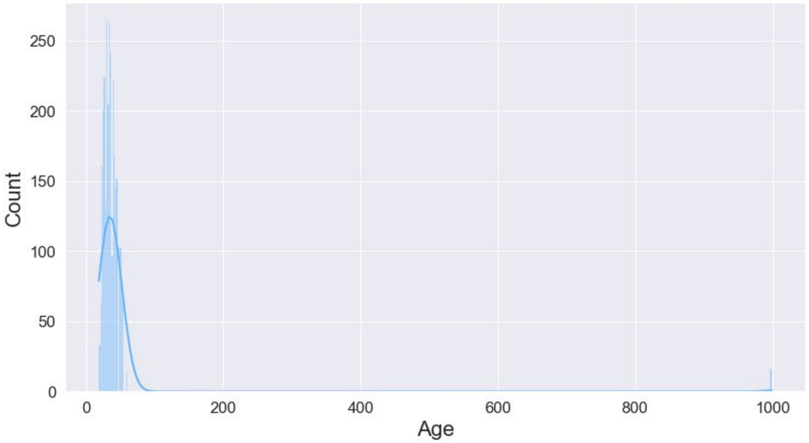
# NUMERICAL VARIABLES

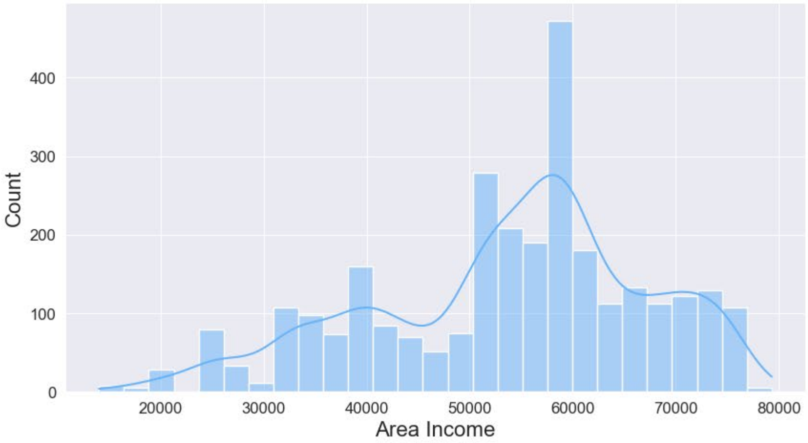


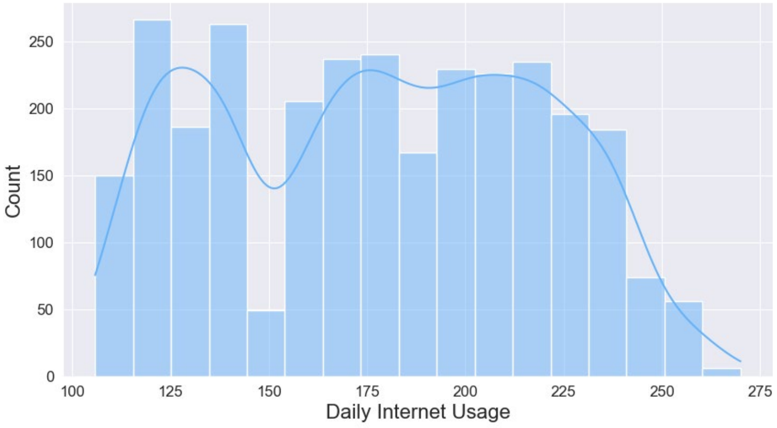


## SKEWNESS

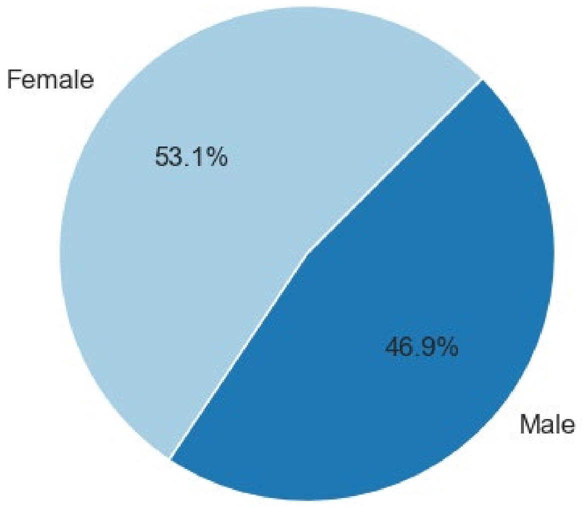
- Timestamp_Combined: **0.06**
- Daily Time Spend on Site: **0.09**
- Age: **13.17**
- Area Income: **-0.51**
- Daily Internet Usage: **-0.01**



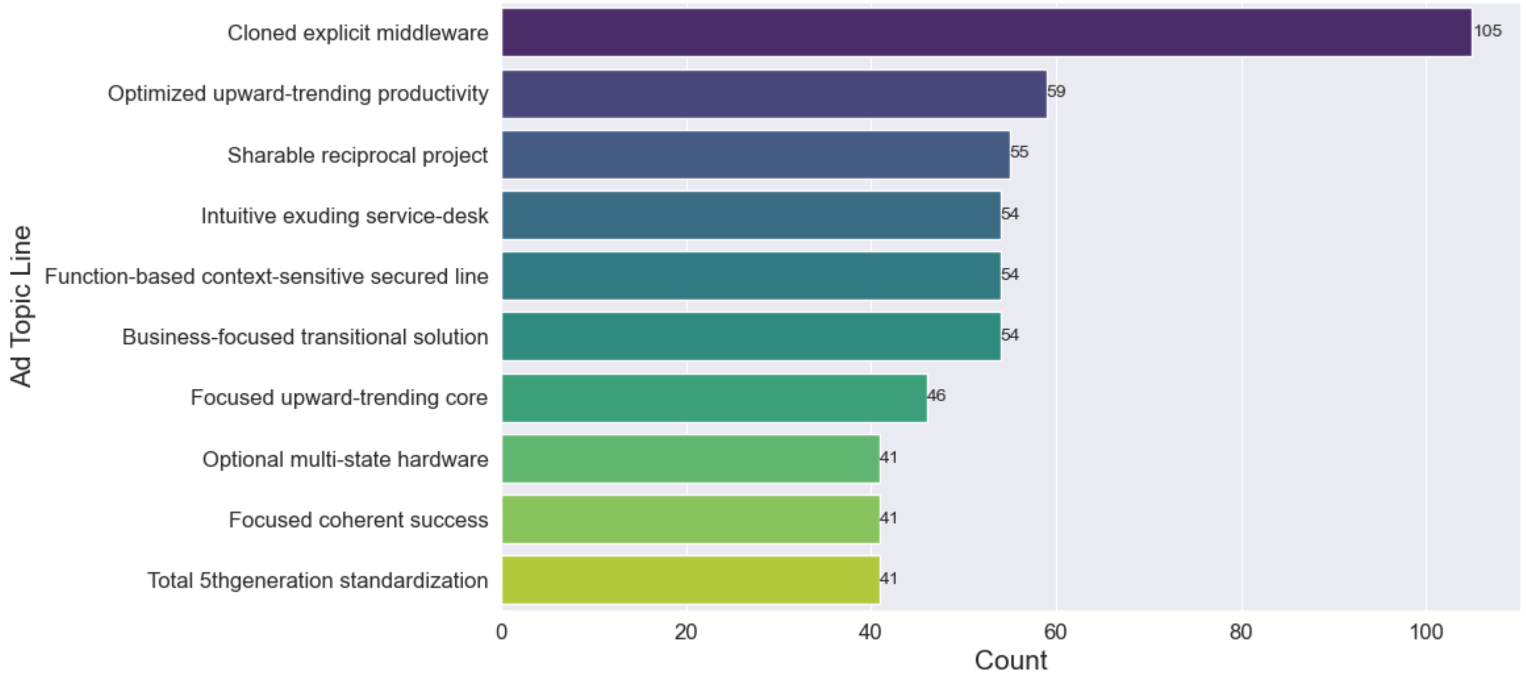




# CATEGORICAL VARIABLES

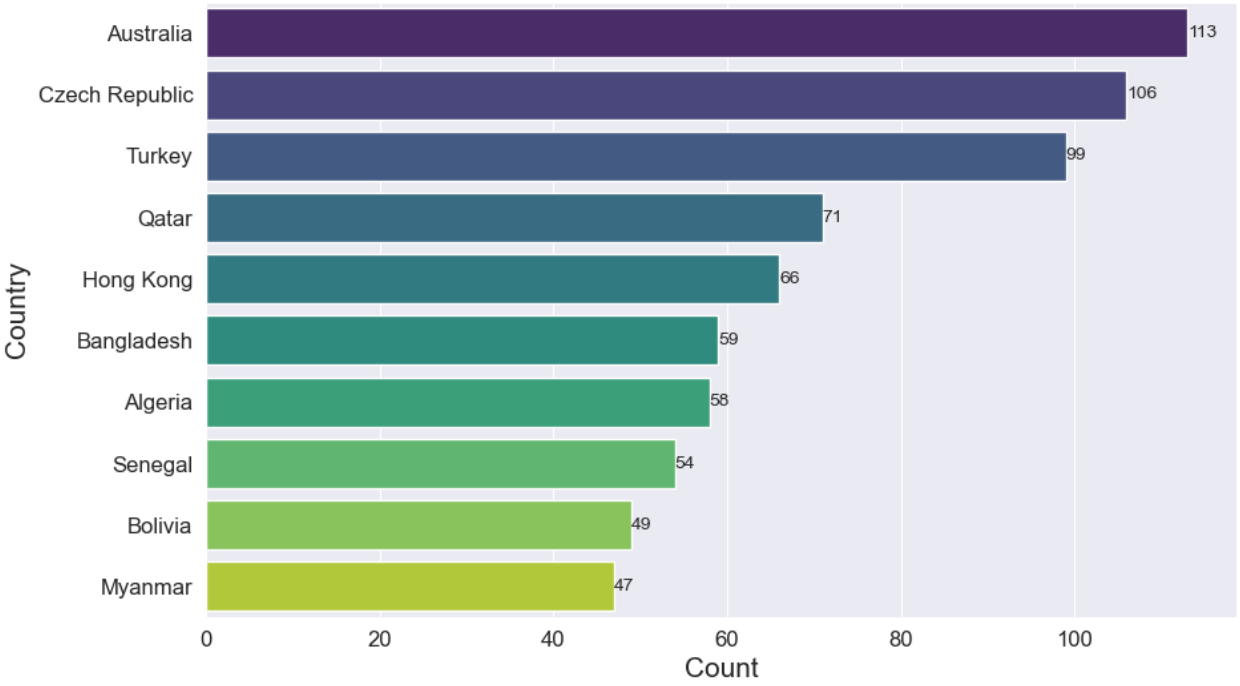


**Gender**

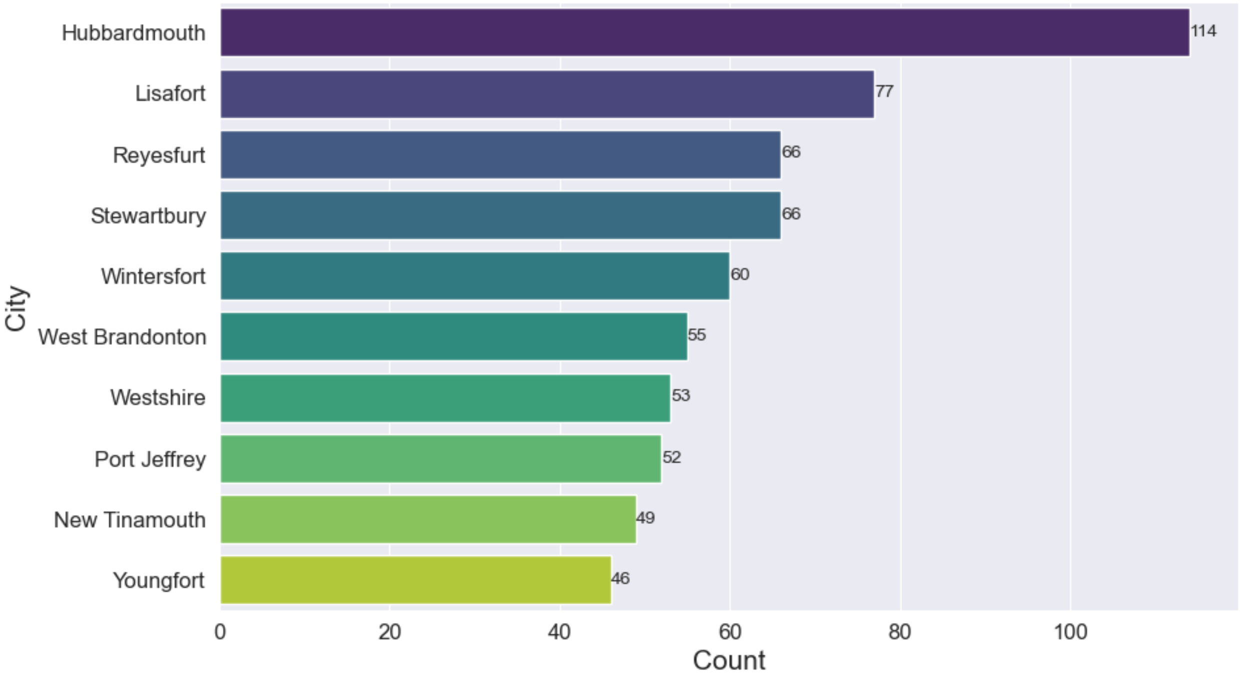


**Ad Topic Line** (417 levels)

# CATEGORICAL VARIABLES



**Country** (186 levels)



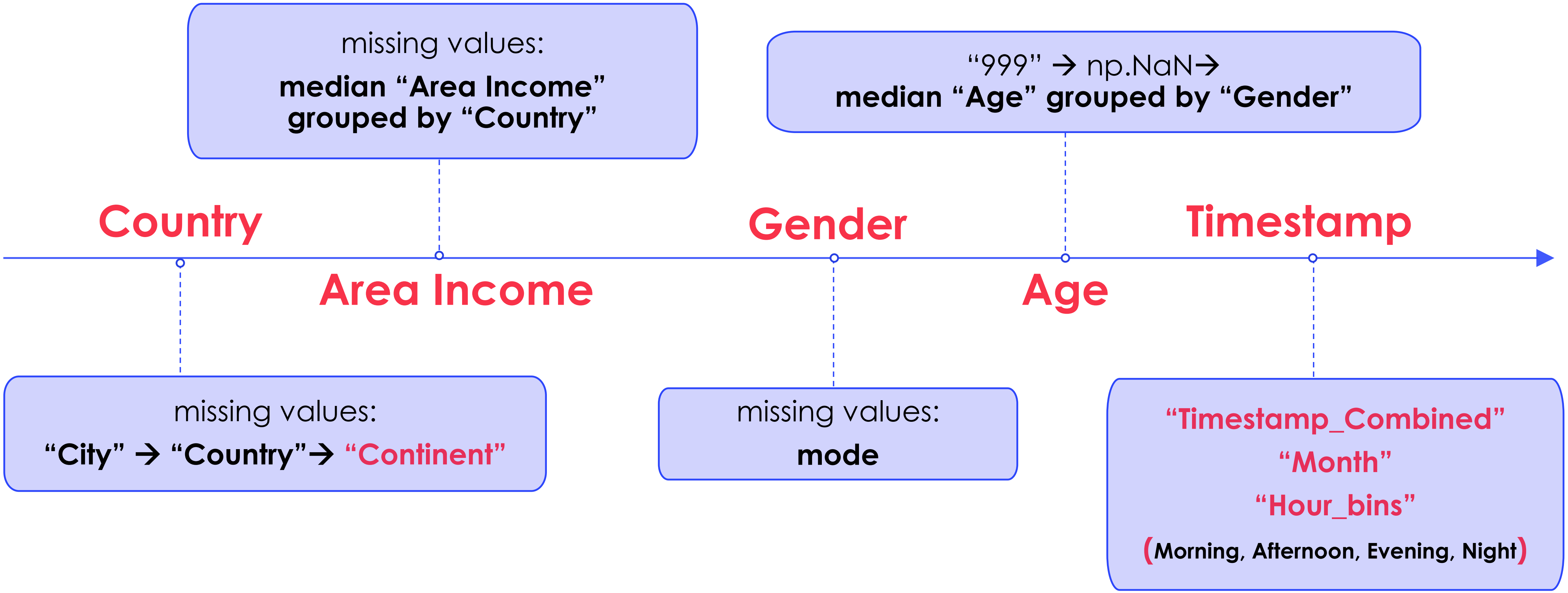
**City** (407 levels)

# FEATURE ENGINEERING

# KEY STEPS

- Handle missing values
- Handle Outliers
- **Feature Creation** (Datetime: Month, Hour)
- **Feature Transformation** (Segmentation)
- **Feature Selection**
  - Statistical Tests (Chi-Square, T-test etc.)
  - Correlation, Multicollinearity (VIF)
- **One-Hot Encoding** for nominal categories (unordered)
- **Feature Scaling and Normalization** (Linear Models)

# FEATURE ENGINEERING



Country

missing values:
“City” → “Country”→ “Continent”

Area Income

missing values:
median “Area Income” grouped by “Country”

Gender

missing values:
mode

Age

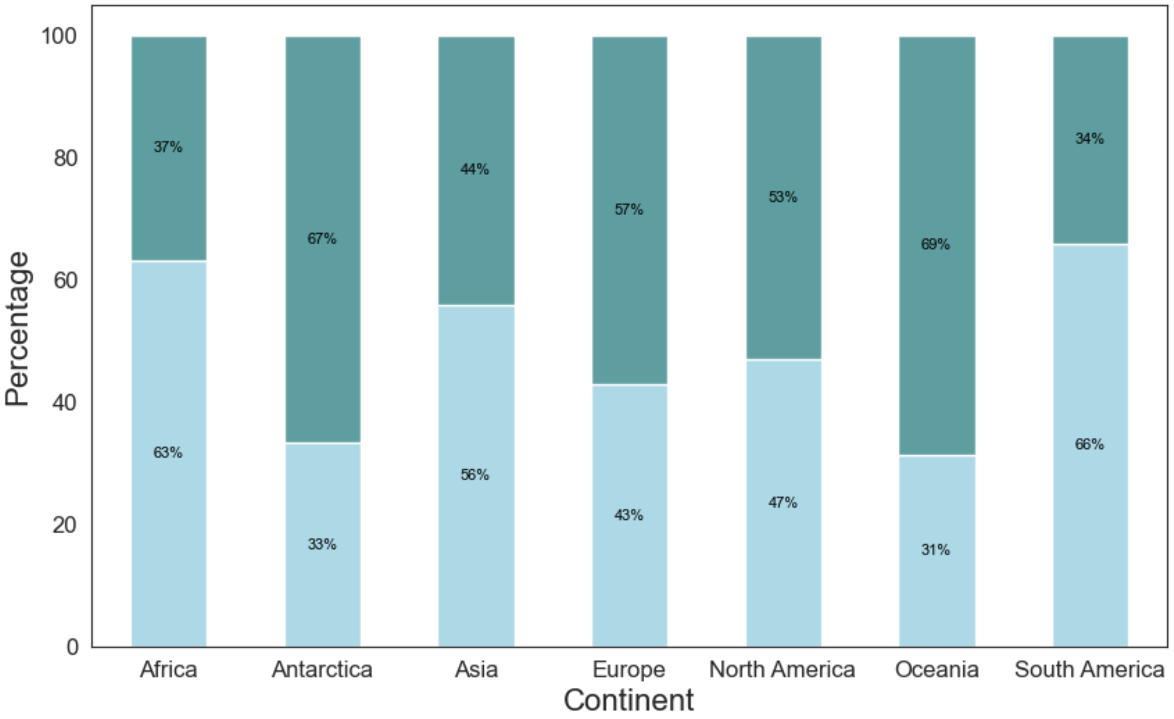
“999” → np.NaN→
median “Age” grouped by “Gender”

Timestamp

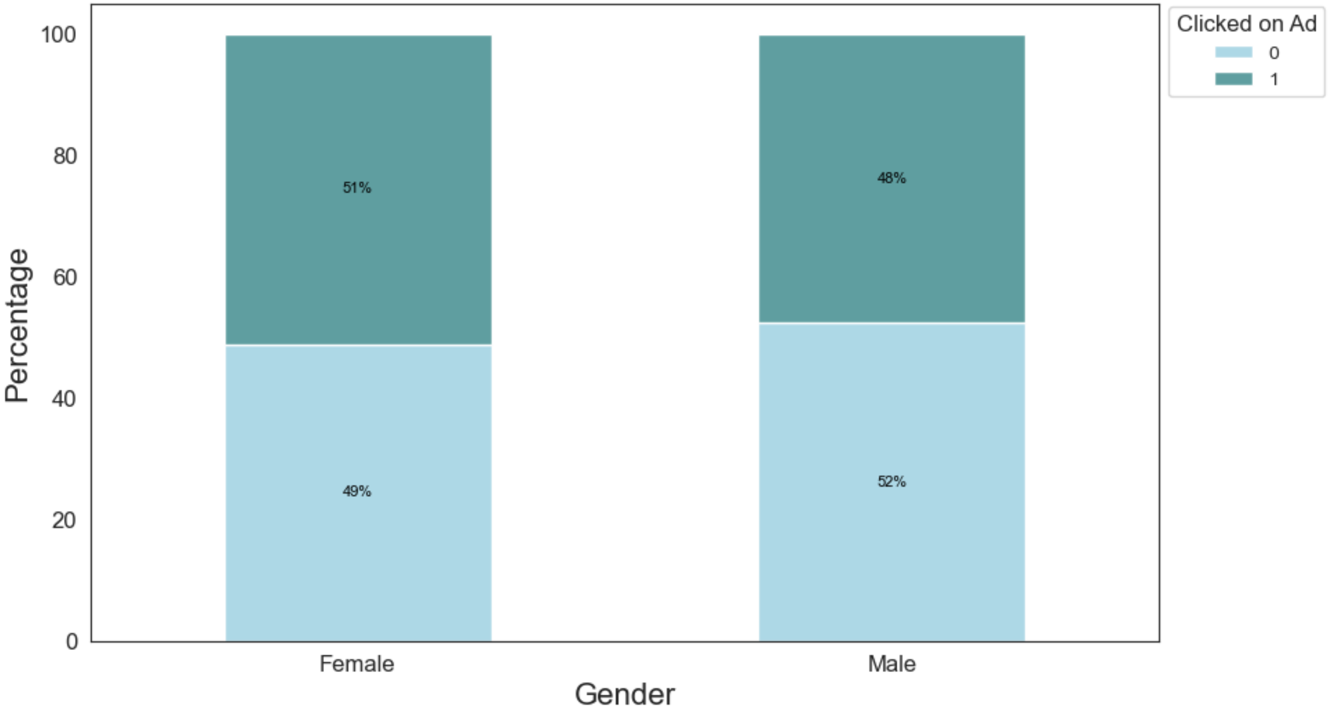
“Timestamp_Combined”
“Month”
“Hour_bins”
(Morning, Afternoon, Evening, Night)

# “CLICKED ON AD” VS CATEGORICAL VARIABLES

## Chi-Square test



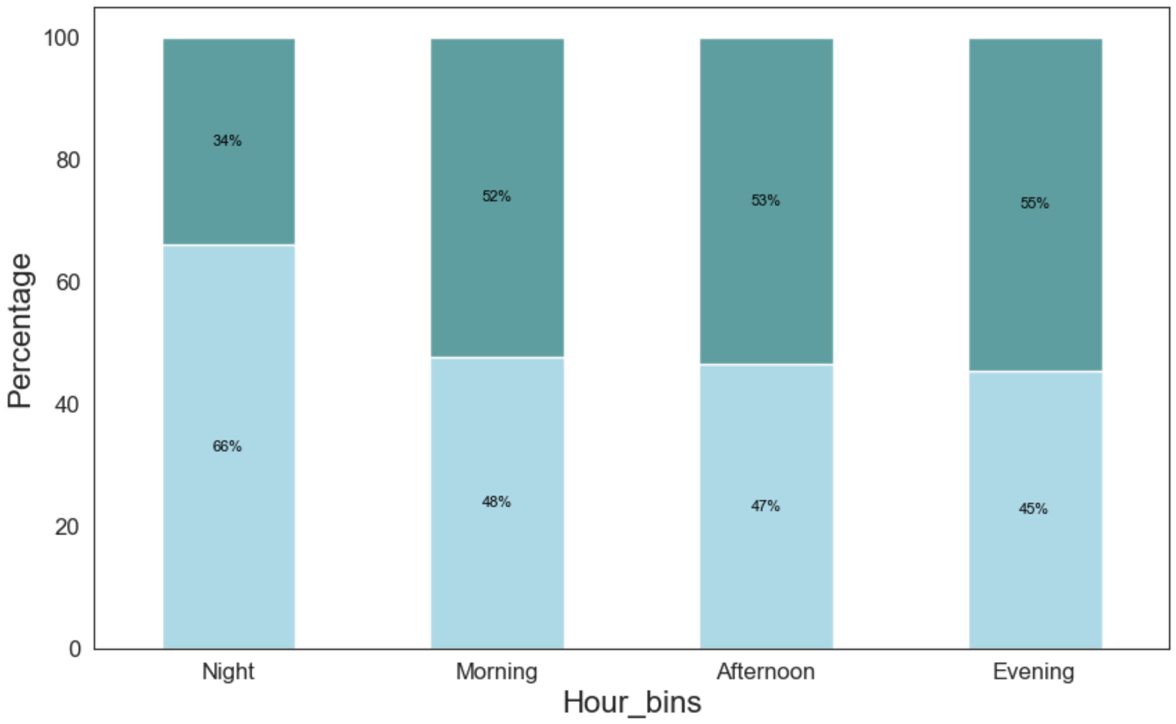
**Continent - Association**



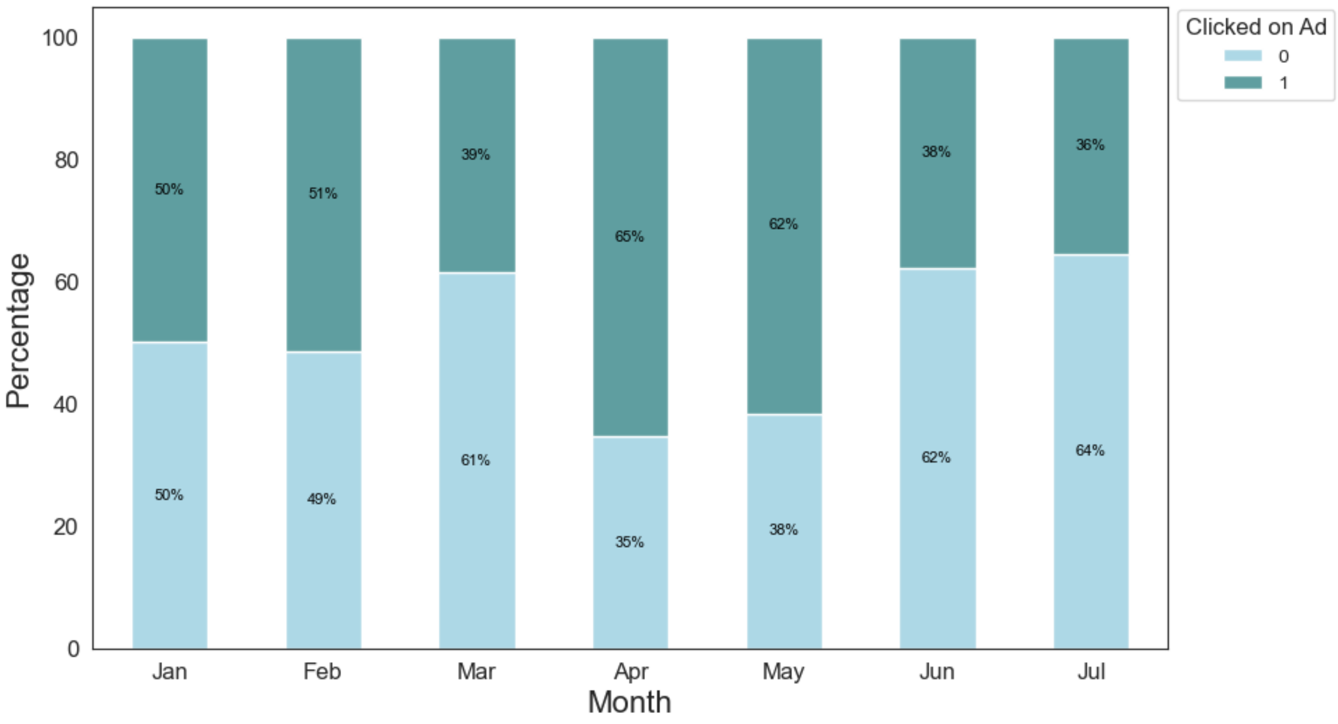
**Gender - No Association**

# “CLICKED ON AD” VS CATEGORICAL VARIABLES

## Chi-Square test



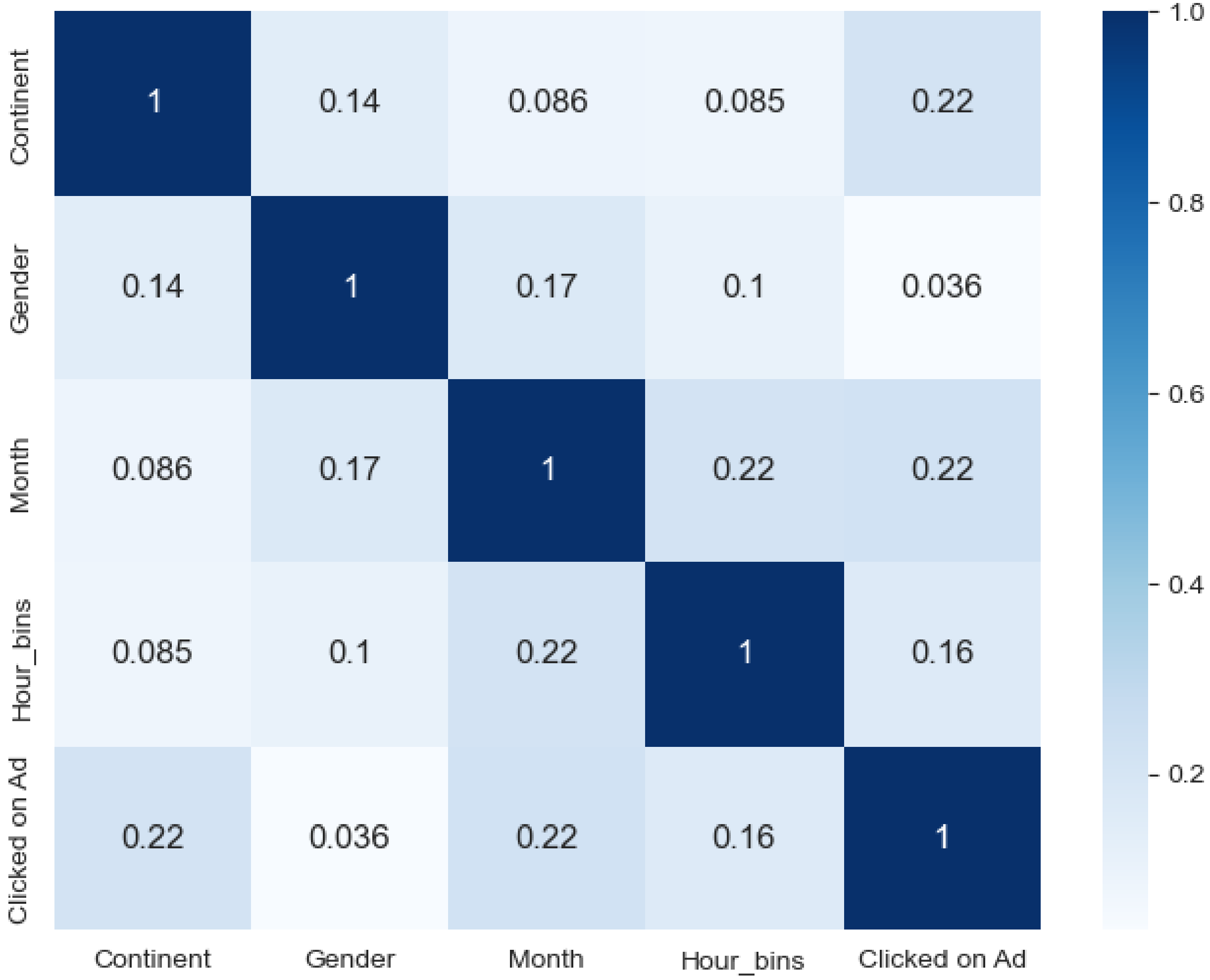
Hour_bins - Association



Month - Association

# "CLICKED ON AD" VS CATEGORICAL VARIABLES

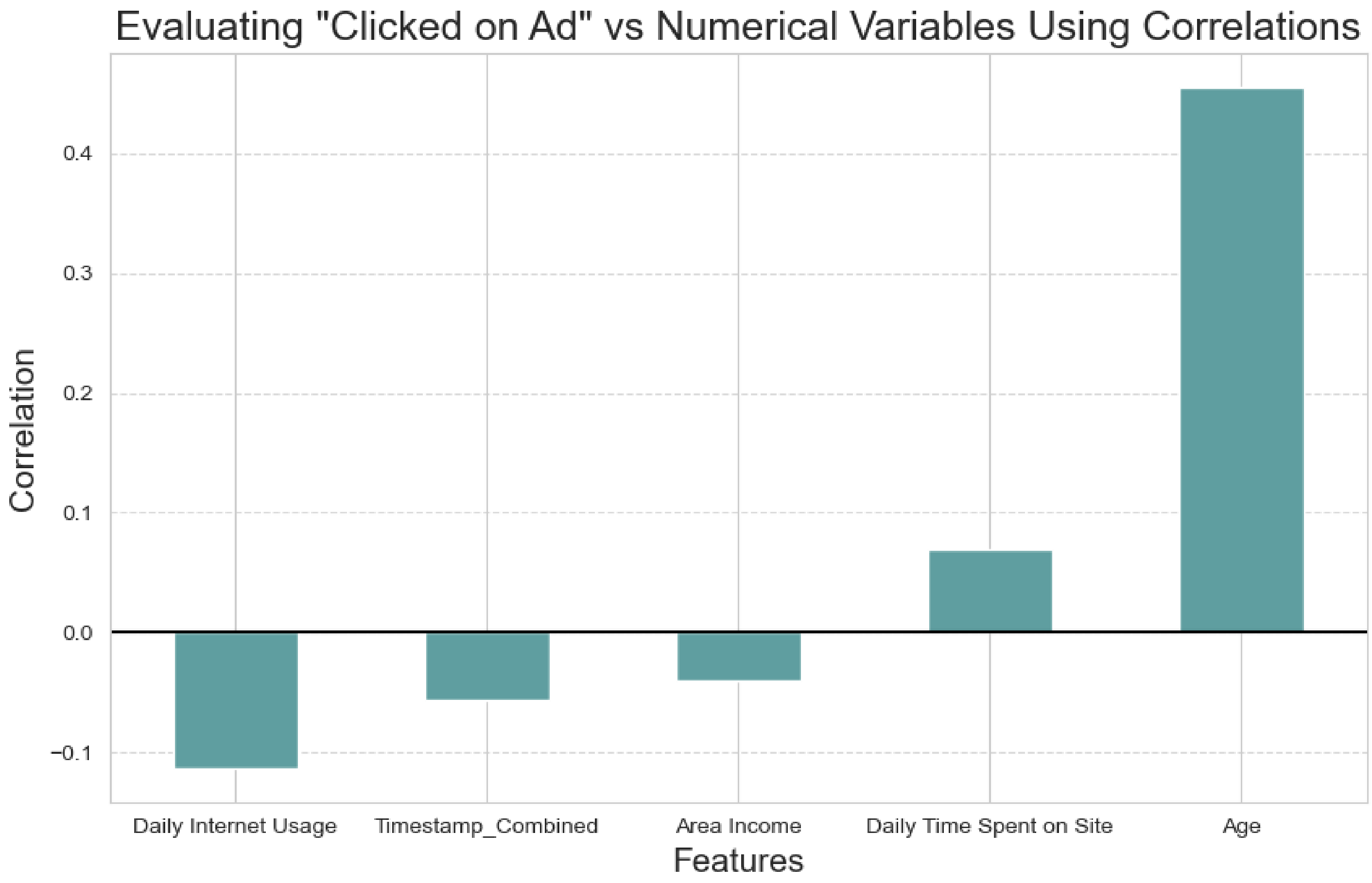
## Cramer's V



|  | Continent | Gender | Month | Hour_bins | Clicked on Ad |
|---|---|---|---|---|---|
| Continent | 1 | 0.14 | 0.086 | 0.085 | 0.22 |
| Gender | 0.14 | 1 | 0.17 | 0.1 | 0.036 |
| Month | 0.086 | 0.17 | 1 | 0.22 | 0.22 |
| Hour_bins | 0.085 | 0.1 | 0.22 | 1 | 0.16 |
| Clicked on Ad | 0.22 | 0.036 | 0.22 | 0.16 | 1 |

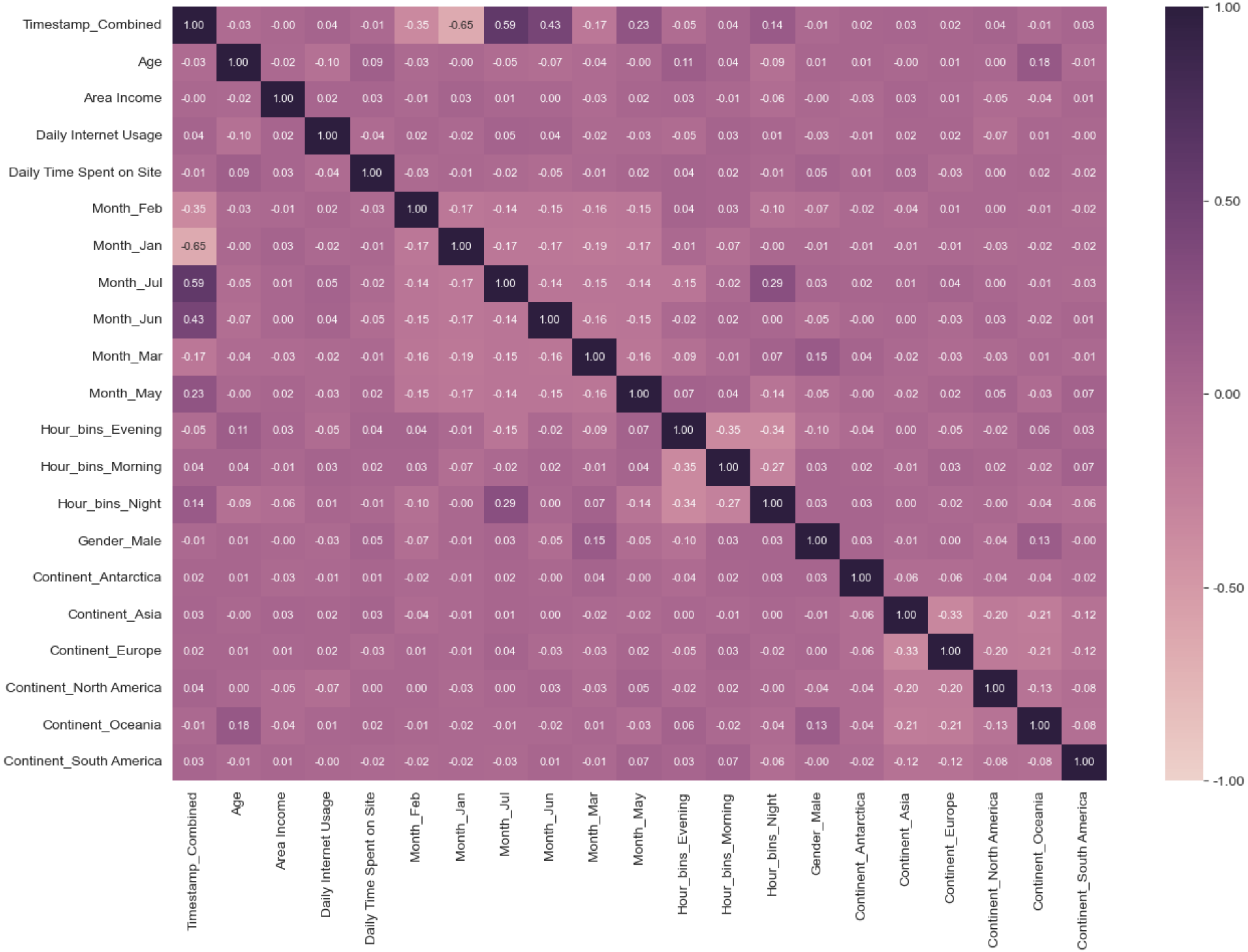
# "CLICKED ON AD" VS NUMERICAL VARIABLES

## T-test



"Age" : Strongest Positive Correlation

# COLLINEARITY



| | Timestamp_Combined | Age | Area Income | Daily Internet Usage | Daily Time Spent on Site | Month_Feb | Month_Jan | Month_Jul | Month_Jun | Month_Mar | Month_May | Hour_bins_Evening | Hour_bins_Morning | Hour_bins_Night | Gender_Male | Continent_Antarctica | Continent_Asia | Continent_Europe | Continent_North America | Continent_Oceania | Continent_South America |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Timestamp_Combined | 1.00 | -0.03 | -0.00 | 0.04 | -0.01 | -0.35 | -0.65 | 0.59 | 0.43 | -0.17 | 0.23 | -0.05 | 0.04 | 0.14 | -0.01 | 0.02 | 0.03 | 0.02 | 0.04 | -0.01 | 0.03 |
| Age | -0.03 | 1.00 | -0.02 | -0.10 | 0.09 | -0.03 | -0.00 | -0.05 | -0.07 | -0.04 | -0.00 | 0.11 | 0.04 | -0.09 | 0.01 | 0.01 | -0.00 | 0.01 | 0.00 | 0.18 | -0.01 |
| Area Income | -0.00 | -0.02 | 1.00 | 0.02 | 0.03 | -0.01 | 0.03 | 0.01 | 0.00 | -0.03 | 0.02 | 0.03 | -0.01 | -0.06 | -0.00 | -0.03 | 0.03 | 0.01 | -0.05 | -0.04 | 0.01 |
| Daily Internet Usage | 0.04 | -0.10 | 0.02 | 1.00 | -0.04 | 0.02 | -0.02 | 0.05 | 0.04 | -0.02 | -0.03 | -0.05 | 0.03 | 0.01 | -0.03 | -0.01 | 0.02 | 0.02 | -0.07 | 0.01 | -0.00 |
| Daily Time Spent on Site | -0.01 | 0.09 | 0.03 | -0.04 | 1.00 | -0.03 | -0.01 | -0.02 | -0.05 | -0.01 | 0.02 | 0.04 | 0.02 | -0.01 | 0.05 | 0.01 | 0.03 | -0.03 | 0.00 | 0.02 | -0.02 |
| Month_Feb | -0.35 | -0.03 | -0.01 | 0.02 | -0.03 | 1.00 | -0.17 | -0.14 | -0.15 | -0.16 | -0.15 | 0.04 | 0.03 | -0.10 | -0.07 | -0.02 | -0.04 | 0.01 | 0.00 | -0.01 | -0.02 |
| Month_Jan | -0.65 | -0.00 | 0.03 | -0.02 | -0.01 | -0.17 | 1.00 | -0.17 | -0.17 | -0.19 | -0.17 | -0.01 | -0.07 | -0.00 | -0.01 | -0.01 | -0.01 | -0.01 | -0.03 | -0.02 | -0.02 |
| Month_Jul | 0.59 | -0.05 | 0.01 | 0.05 | -0.02 | -0.14 | -0.17 | 1.00 | -0.14 | -0.15 | -0.14 | -0.15 | -0.02 | 0.29 | 0.03 | 0.02 | 0.01 | 0.04 | 0.00 | -0.01 | -0.03 |
| Month_Jun | 0.43 | -0.07 | 0.00 | 0.04 | -0.05 | -0.15 | -0.17 | -0.14 | 1.00 | -0.16 | -0.15 | -0.02 | 0.02 | 0.00 | -0.05 | -0.00 | 0.00 | -0.03 | 0.03 | -0.02 | 0.01 |
| Month_Mar | -0.17 | -0.04 | -0.03 | -0.02 | -0.01 | -0.16 | -0.19 | -0.15 | -0.16 | 1.00 | -0.16 | -0.09 | -0.01 | 0.07 | 0.15 | 0.04 | -0.02 | -0.03 | -0.03 | 0.01 | -0.01 |
| Month_May | 0.23 | -0.00 | 0.02 | -0.03 | 0.02 | -0.15 | -0.17 | -0.14 | -0.15 | -0.16 | 1.00 | 0.07 | 0.04 | -0.14 | -0.05 | -0.00 | -0.02 | 0.02 | 0.05 | -0.03 | 0.07 |
| Hour_bins_Evening | -0.05 | 0.11 | 0.03 | -0.05 | 0.04 | 0.04 | -0.01 | -0.15 | -0.02 | -0.09 | 0.07 | 1.00 | -0.35 | -0.34 | -0.10 | -0.04 | 0.00 | -0.05 | -0.02 | 0.06 | 0.03 |
| Hour_bins_Morning | 0.04 | 0.04 | -0.01 | 0.03 | 0.02 | 0.03 | -0.07 | -0.02 | 0.02 | -0.01 | 0.04 | -0.35 | 1.00 | -0.27 | 0.03 | 0.02 | -0.01 | 0.03 | 0.02 | -0.02 | 0.07 |
| Hour_bins_Night | 0.14 | -0.09 | -0.06 | 0.01 | -0.01 | -0.10 | -0.00 | 0.29 | 0.00 | 0.07 | -0.14 | -0.34 | -0.27 | 1.00 | 0.03 | 0.03 | 0.00 | -0.02 | -0.00 | -0.04 | -0.06 |
| Gender_Male | -0.01 | 0.01 | -0.00 | -0.03 | 0.05 | -0.07 | -0.01 | 0.03 | -0.05 | 0.15 | -0.05 | -0.10 | 0.03 | 0.03 | 1.00 | 0.03 | -0.01 | 0.00 | -0.04 | 0.13 | -0.00 |
| Continent_Antarctica | 0.02 | 0.01 | -0.03 | -0.01 | 0.01 | -0.02 | -0.01 | 0.02 | -0.00 | 0.04 | -0.00 | -0.04 | 0.02 | 0.03 | 0.03 | 1.00 | -0.06 | -0.06 | -0.04 | -0.04 | -0.02 |
| Continent_Asia | 0.03 | -0.00 | 0.03 | 0.02 | 0.03 | -0.04 | -0.01 | 0.01 | 0.00 | -0.02 | -0.02 | 0.00 | -0.01 | 0.00 | -0.01 | -0.06 | 1.00 | -0.33 | -0.20 | -0.21 | -0.12 |
| Continent_Europe | 0.02 | 0.01 | 0.01 | 0.02 | -0.03 | 0.01 | -0.01 | 0.04 | -0.03 | -0.03 | 0.02 | -0.05 | 0.03 | -0.02 | 0.00 | -0.06 | -0.33 | 1.00 | -0.20 | -0.21 | -0.12 |
| Continent_North America | 0.04 | 0.00 | -0.05 | -0.07 | 0.00 | 0.00 | -0.03 | 0.00 | 0.03 | -0.03 | 0.05 | -0.02 | 0.02 | -0.00 | -0.04 | -0.04 | -0.20 | -0.20 | 1.00 | -0.13 | -0.08 |
| Continent_Oceania | -0.01 | 0.18 | -0.04 | 0.01 | 0.02 | -0.01 | -0.02 | -0.01 | -0.02 | 0.01 | -0.03 | 0.06 | -0.02 | -0.04 | 0.13 | -0.04 | -0.21 | -0.21 | -0.13 | 1.00 | -0.08 |
| Continent_South America | 0.03 | -0.01 | 0.01 | -0.00 | -0.02 | -0.02 | -0.02 | -0.03 | 0.01 | -0.01 | 0.07 | 0.03 | 0.07 | -0.06 | -0.00 | -0.02 | -0.12 | -0.12 | -0.08 | -0.08 | 1.00 |

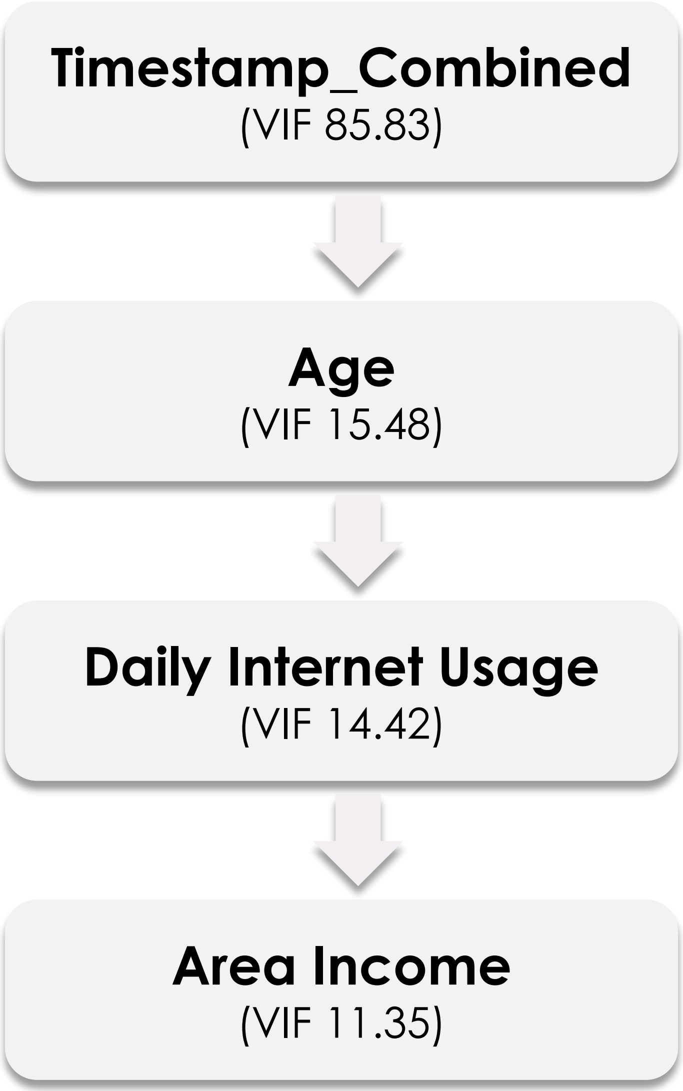
# MULTICOLLINEARITY

## Linear Model

- Remove Features (VIF >10)
    - Unstable Coefficients
    - Reduced Interpretability
- Scale the data

## Tree-Based Models

- Keep All Features
- Scale not necessary



**Timestamp_Combined**
(VIF 85.83)

**Age**
(VIF 15.48)

**Daily Internet Usage**
(VIF 14.42)

**Area Income**
(VIF 11.35)

| Feature | VIF (<=10) |
|---|---|
| Daily Time Spent on Site | 8.04 |
| Month_Feb | 1.49 |
| Month_Jan | 1.66 |
| Month_Jul | 1.65 |
| Month_Jun | 1.49 |
| Month_Mar | 1.64 |
| Month_May | 1.54 |
| Hour_bins_Evening | 2.05 |
| Hour_bins_Morning | 1.77 |
| Hour_bins_Night | 1.86 |
| Gender_Male | 1.96 |
| Continent_Antarctica | 1.05 |
| Continent_Asia | 1.95 |
| Continent_Europe | 1.90 |
| Continent_North America | 1.41 |
| Continent_Oceania | 1.50 |
| Continent_South America | 1.19 |

# MODEL DEVELOPMENT

Logistic Regression

Decision Tree

Random Forest

XGBoost

# ALL MODELS

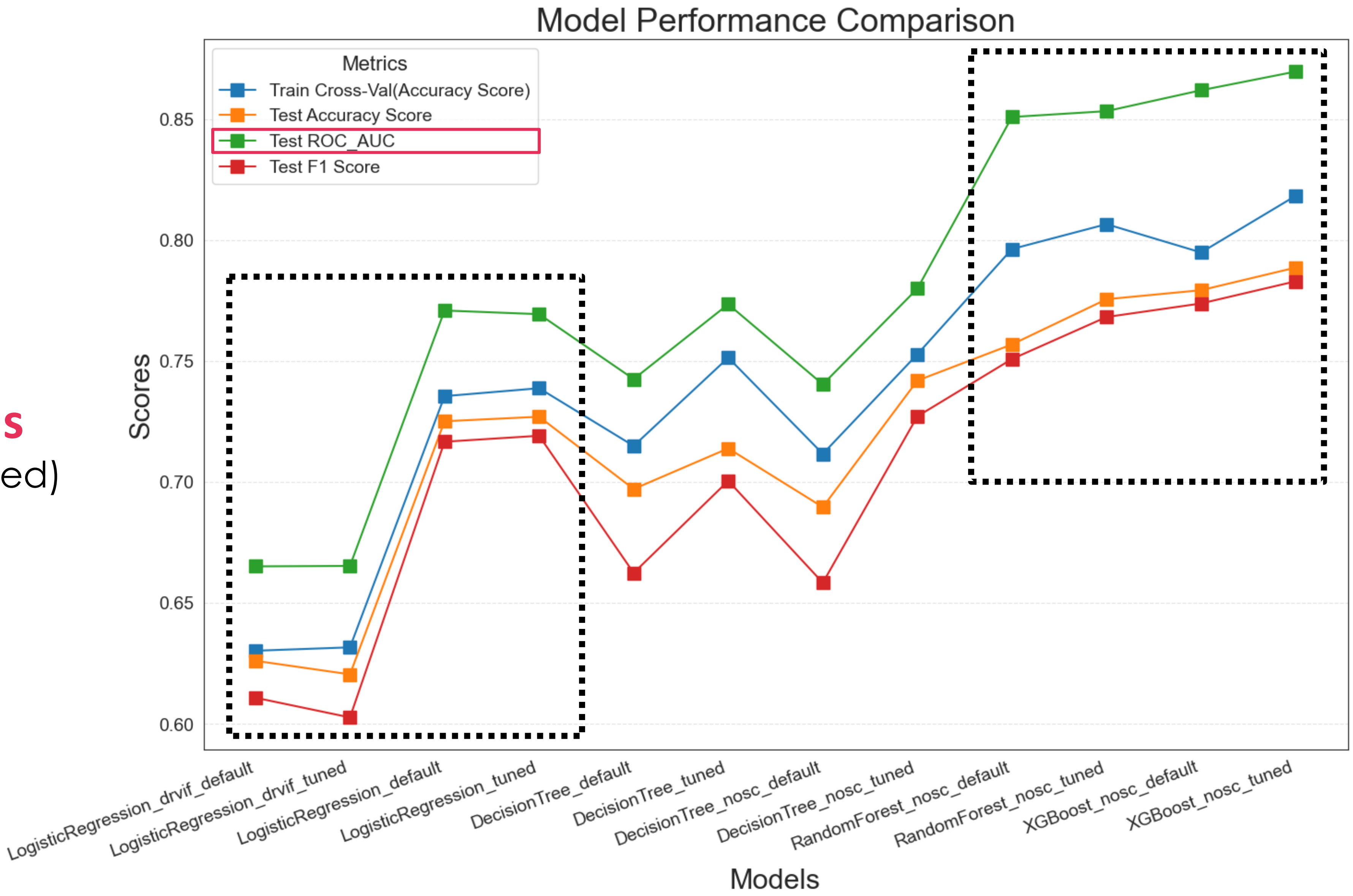
## Model Selection

### Linear Model

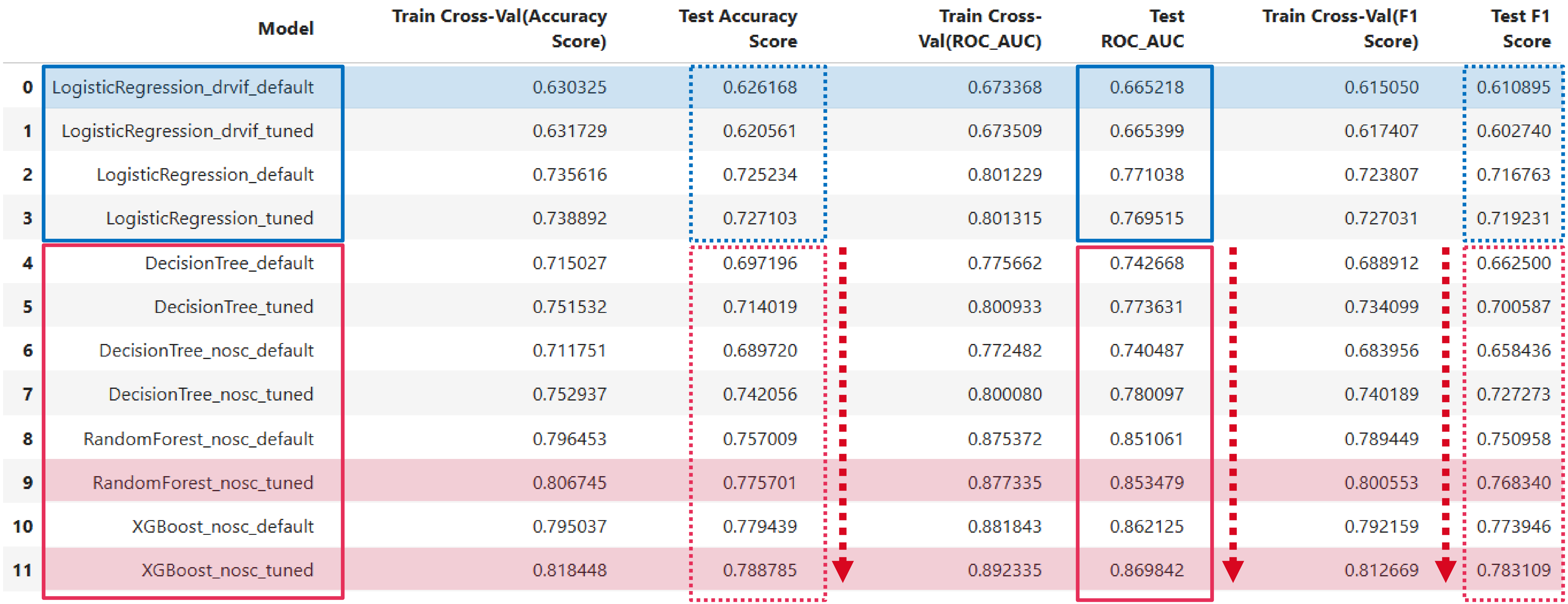
- Logistic Regression (Baseline)

### Tree-Based Models

- Random Forest (Tuned)
- XGBoost (Tuned)



# ALL MODELS

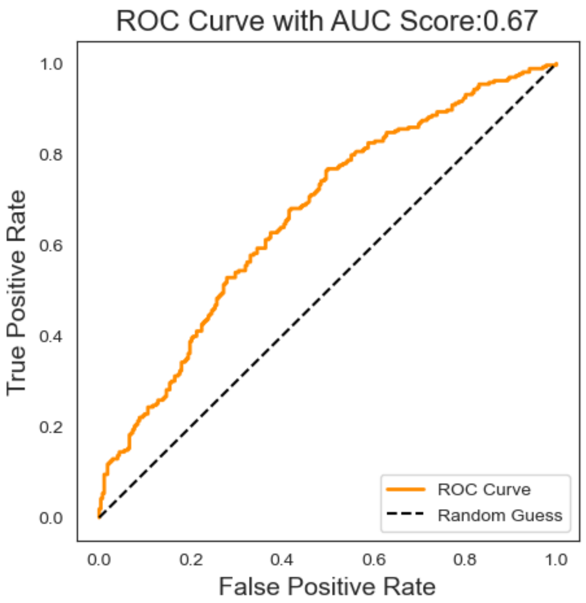


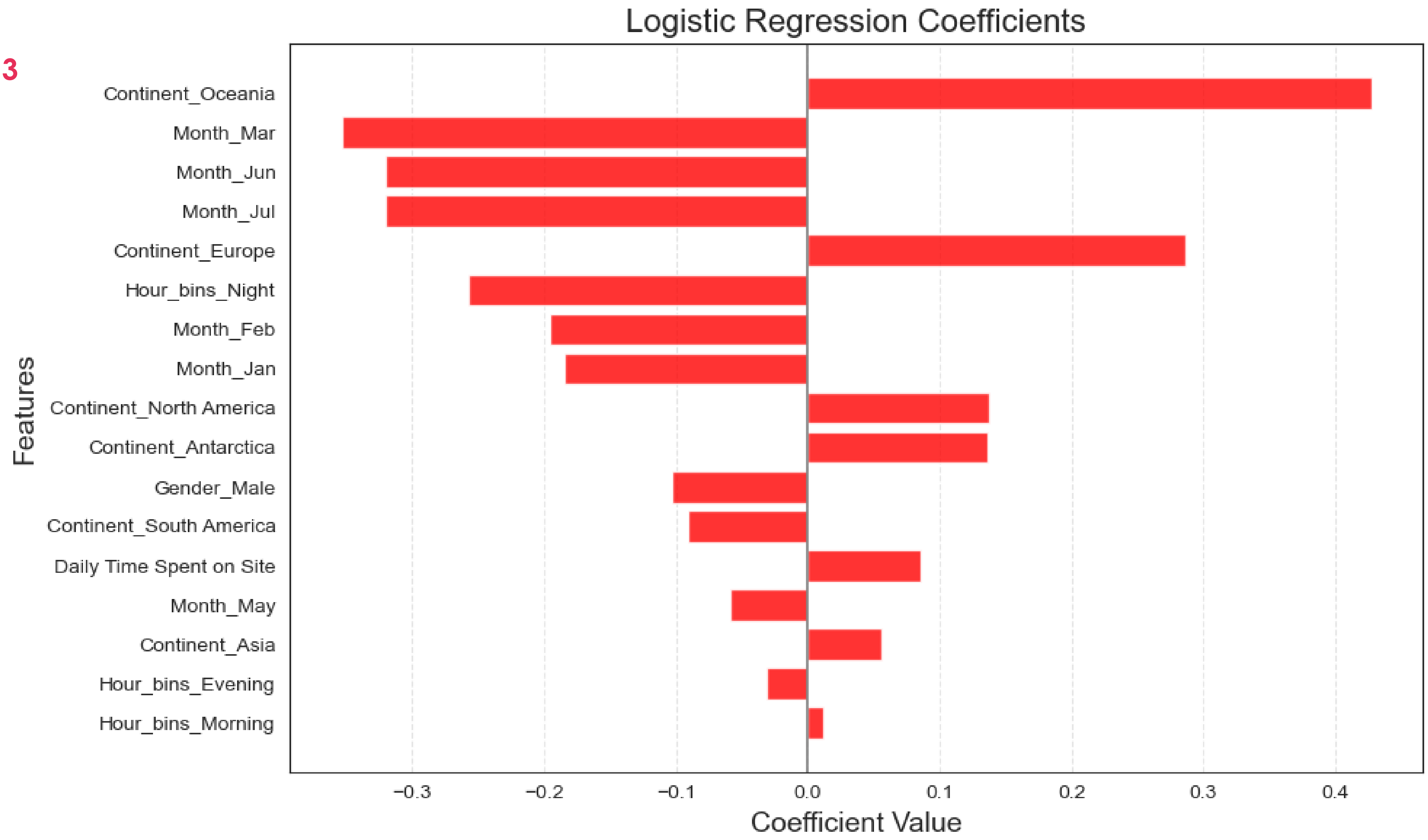
| | Model | Train Cross-Val(Accuracy Score) | Test Accuracy Score | Train Cross-Val(ROC_AUC) | Test ROC_AUC | Train Cross-Val(F1 Score) | Test F1 Score |
|---|---|---|---|---|---|---|---|
| 0 | LogisticRegression_drvif_default | 0.630325 | 0.626168 | 0.673368 | 0.665218 | 0.615050 | 0.610895 |
| 1 | LogisticRegression_drvif_tuned | 0.631729 | 0.620561 | 0.673509 | 0.665399 | 0.617407 | 0.602740 |
| 2 | LogisticRegression_default | 0.735616 | 0.725234 | 0.801229 | 0.771038 | 0.723807 | 0.716763 |
| 3 | LogisticRegression_tuned | 0.738892 | 0.727103 | 0.801315 | 0.769515 | 0.727031 | 0.719231 |
| 4 | DecisionTree_default | 0.715027 | 0.697196 | 0.775662 | 0.742668 | 0.688912 | 0.662500 |
| 5 | DecisionTree_tuned | 0.751532 | 0.714019 | 0.800933 | 0.773631 | 0.734099 | 0.700587 |
| 6 | DecisionTree_nosc_default | 0.711751 | 0.689720 | 0.772482 | 0.740487 | 0.683956 | 0.658436 |
| 7 | DecisionTree_nosc_tuned | 0.752937 | 0.742056 | 0.800080 | 0.780097 | 0.740189 | 0.727273 |
| 8 | RandomForest_nosc_default | 0.796453 | 0.757009 | 0.875372 | 0.851061 | 0.789449 | 0.750958 |
| 9 | RandomForest_nosc_tuned | 0.806745 | 0.775701 | 0.877335 | 0.853479 | 0.800553 | 0.768340 |
| 10 | XGBoost_nosc_default | 0.795037 | 0.779439 | 0.881843 | 0.862125 | 0.792159 | 0.773946 |
| 11 | XGBoost_nosc_tuned | 0.818448 | 0.788785 | 0.892335 | 0.869842 | 0.812669 | 0.783109 |

# LOGISTIC REGRESSION

## Features VIF<=10

- Train Cross-Val Accuracy: **0.63**
- Test Accuracy: **0.62**
- Test ROC_AUC: **0.66**
- Test F1 Score: **0.60**

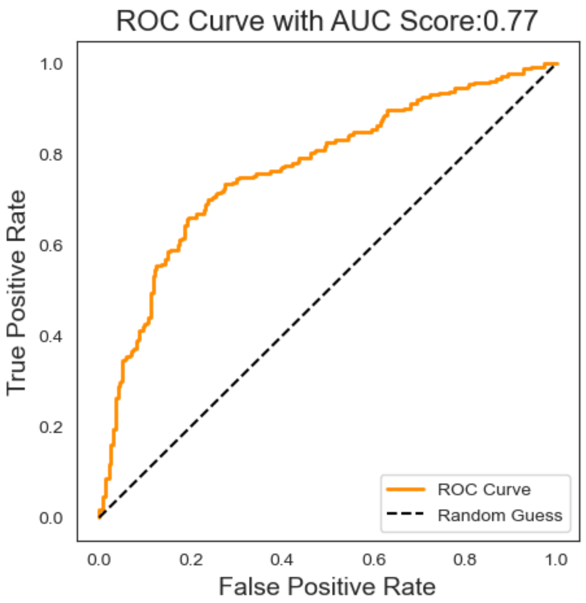


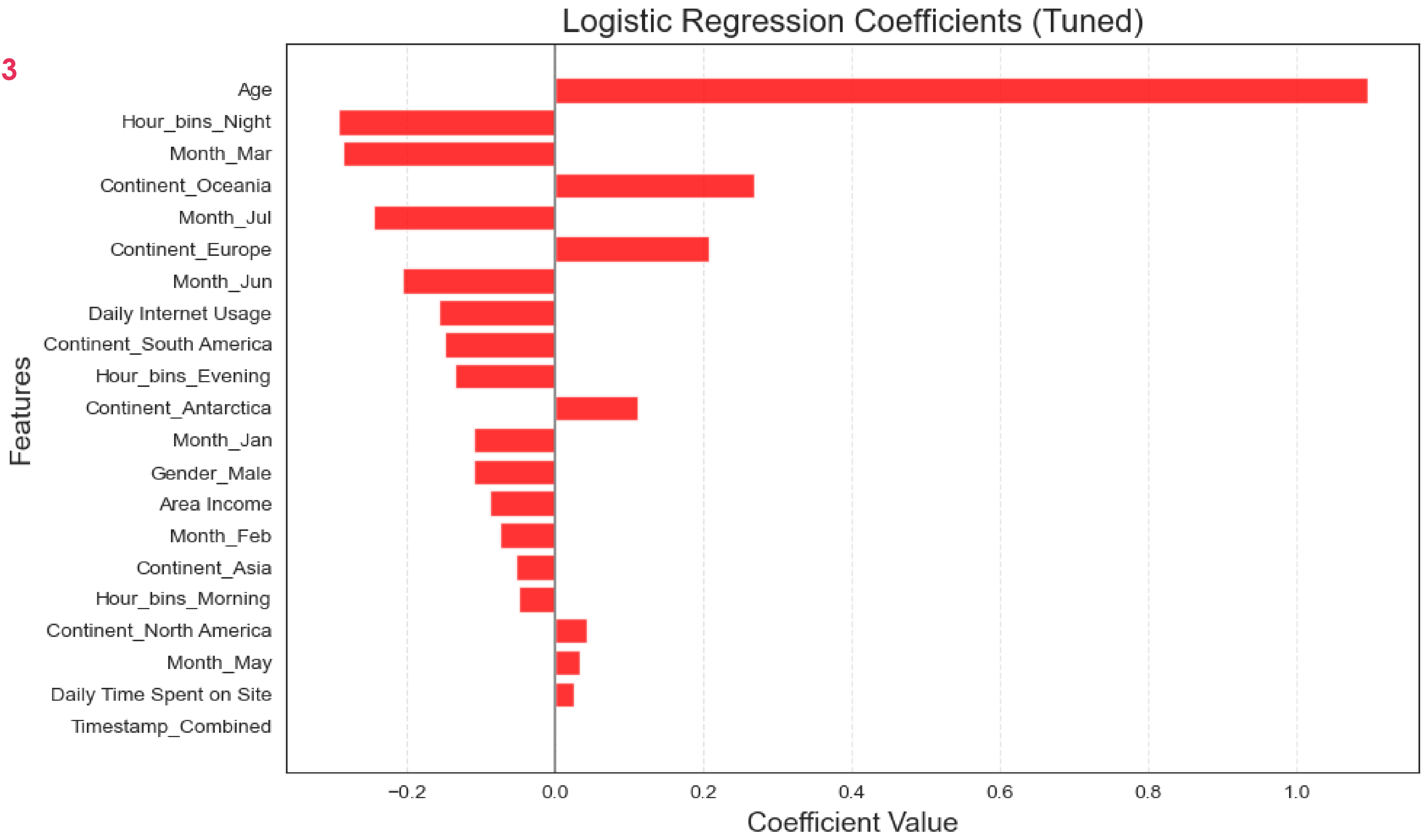


# LOGISTIC REGRESSION

## All Columns

- Train Cross-Val Accuracy: **0.73**
- Test Accuracy: **0.72**
- Test ROC_AUC: **0.76**
- Test F1 Score: **0.71**



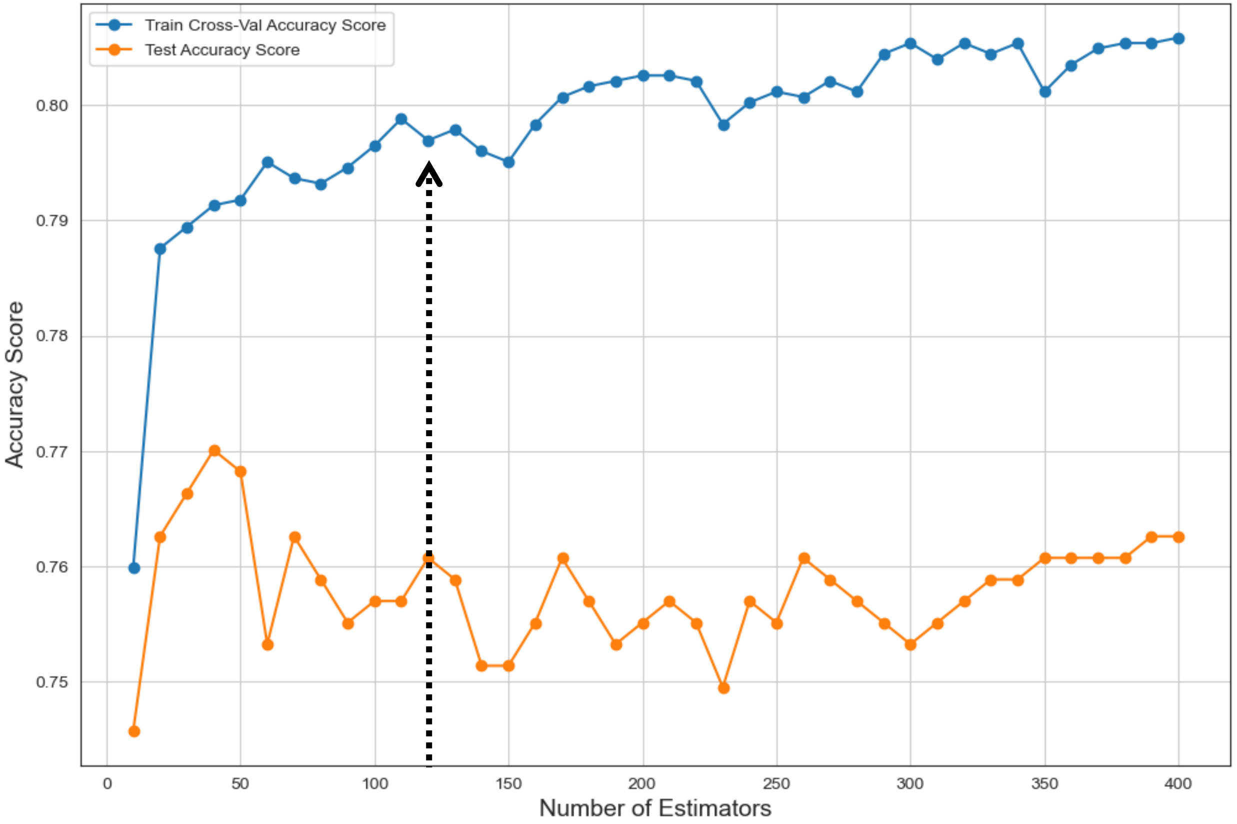


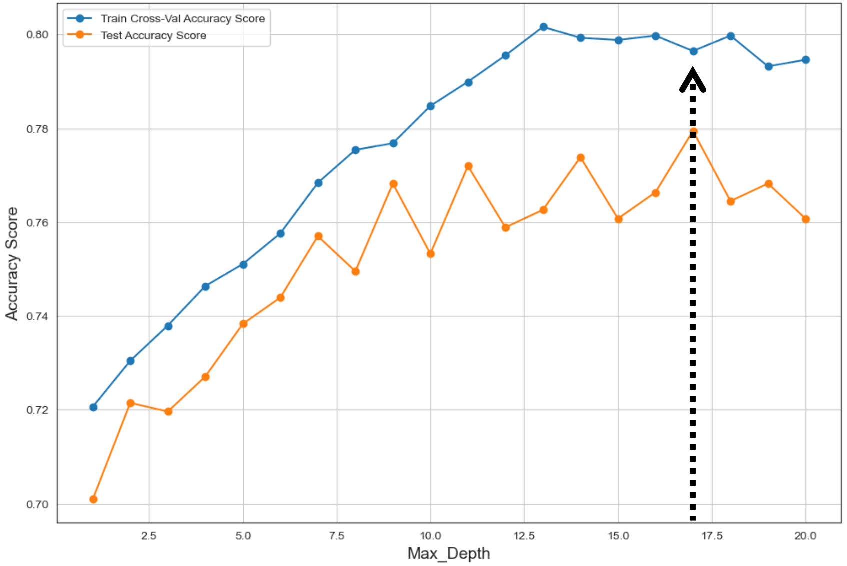
# TREE-BASED MODELS

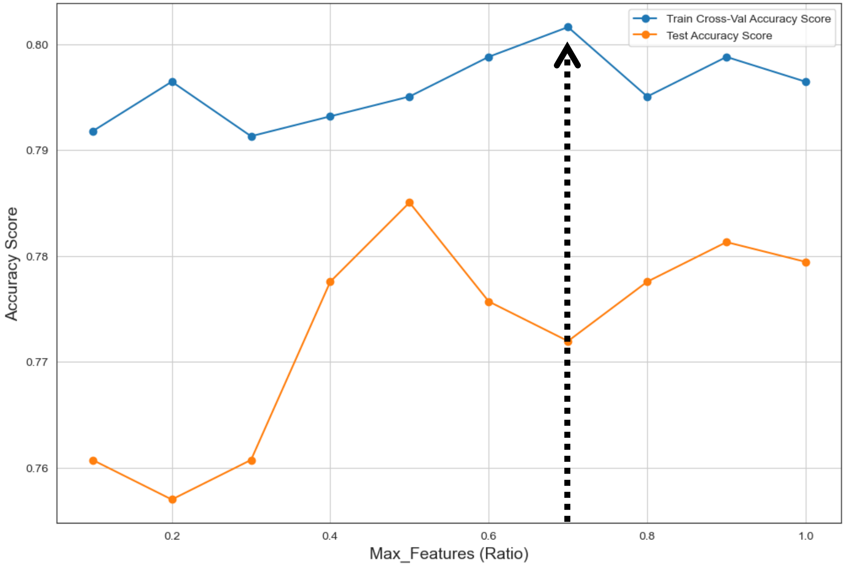
## Random Forest

- Train Cross-Val Accuracy: **0.80**
- Test Accuracy: **0.77**
- Test ROC_AUC: **0.85**
- Test F1 Score: **0.76**

- n_estimators: 120
- max_depth: 17
- max_features: 0.7

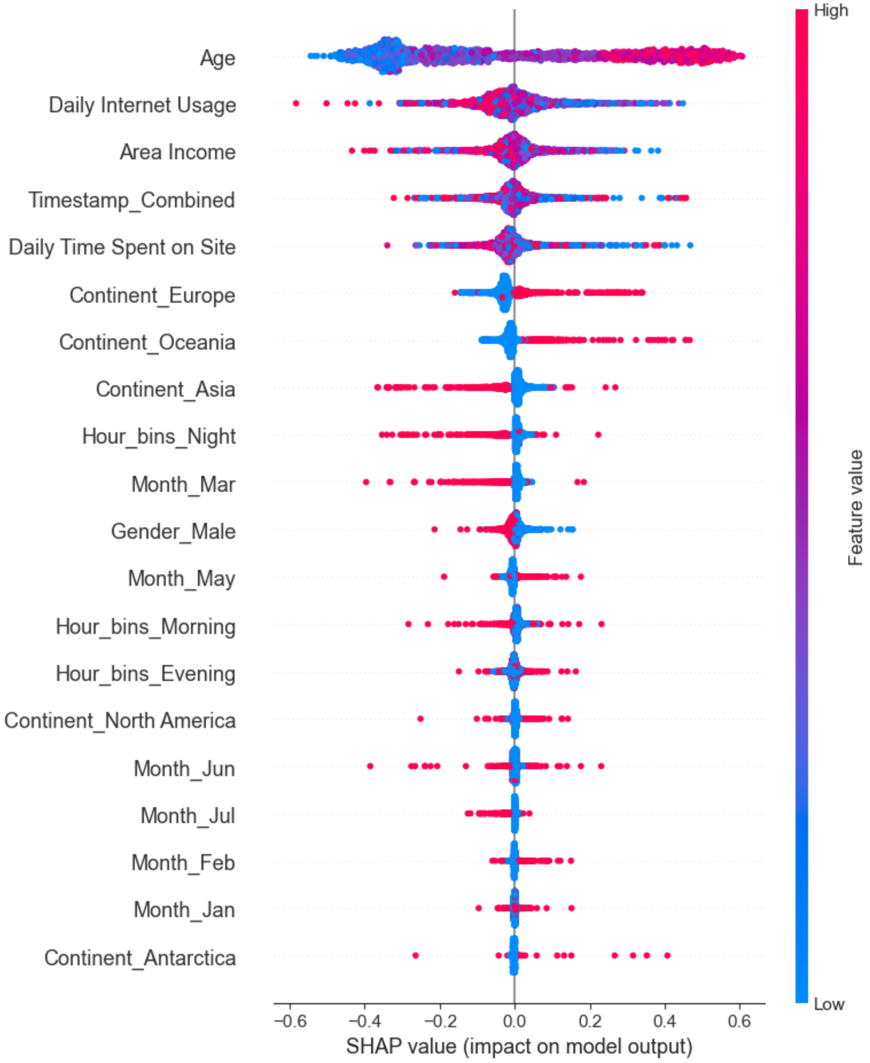


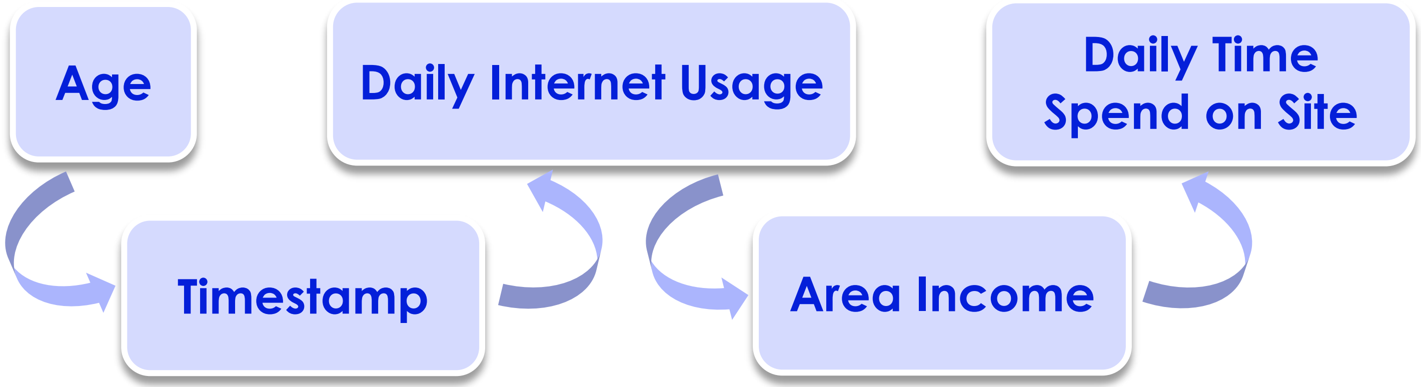




# TREE-BASED MODELS

## Random Forest





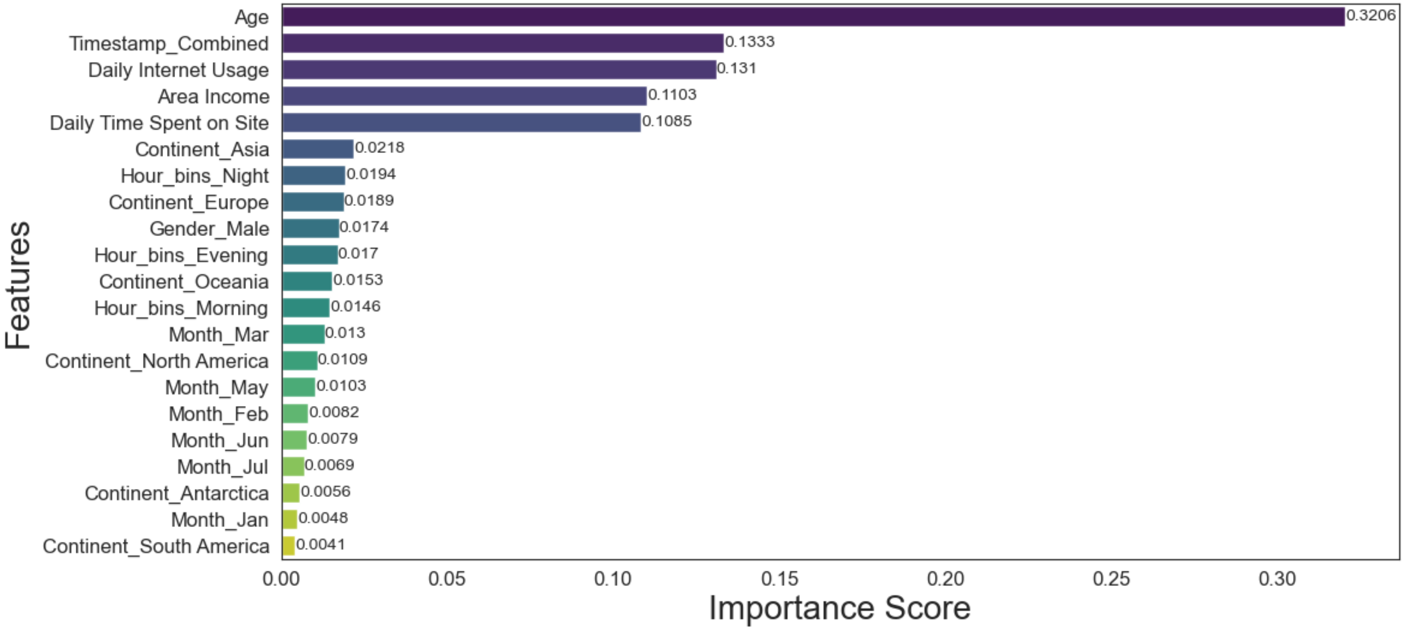
Age

Timestamp

Daily Internet Usage

Area Income

Daily Time Spend on Site

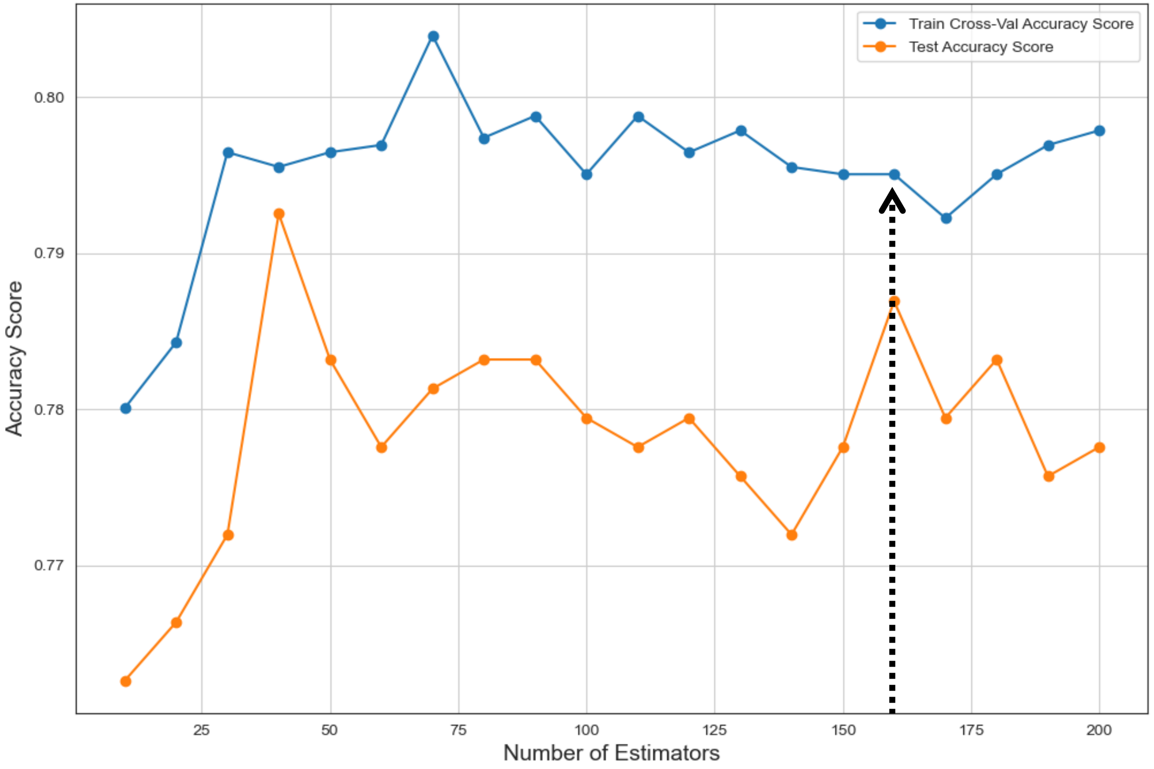


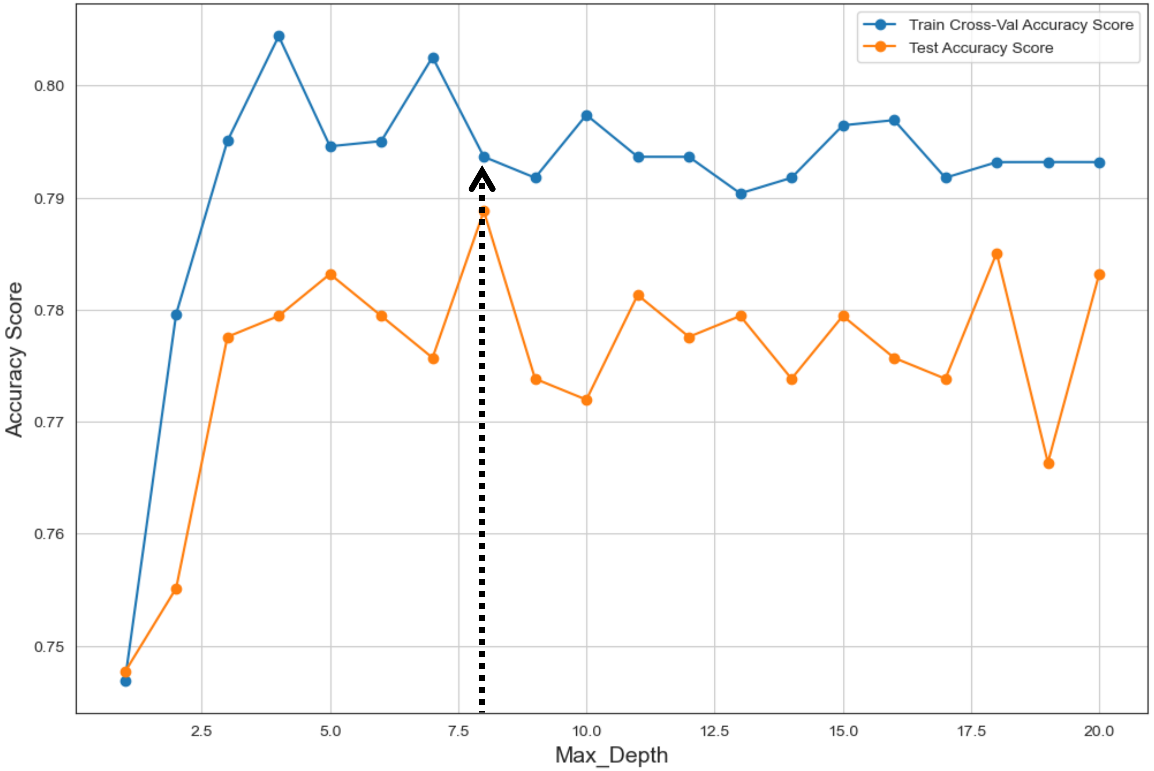
# TREE-BASED MODELS

## XGBoost

- Train Cross-Val Accuracy: **0.81**
- Test Accuracy: **0.78**
- Test ROC_AUC: **0.86**
- Test F1 Score: **0.78**

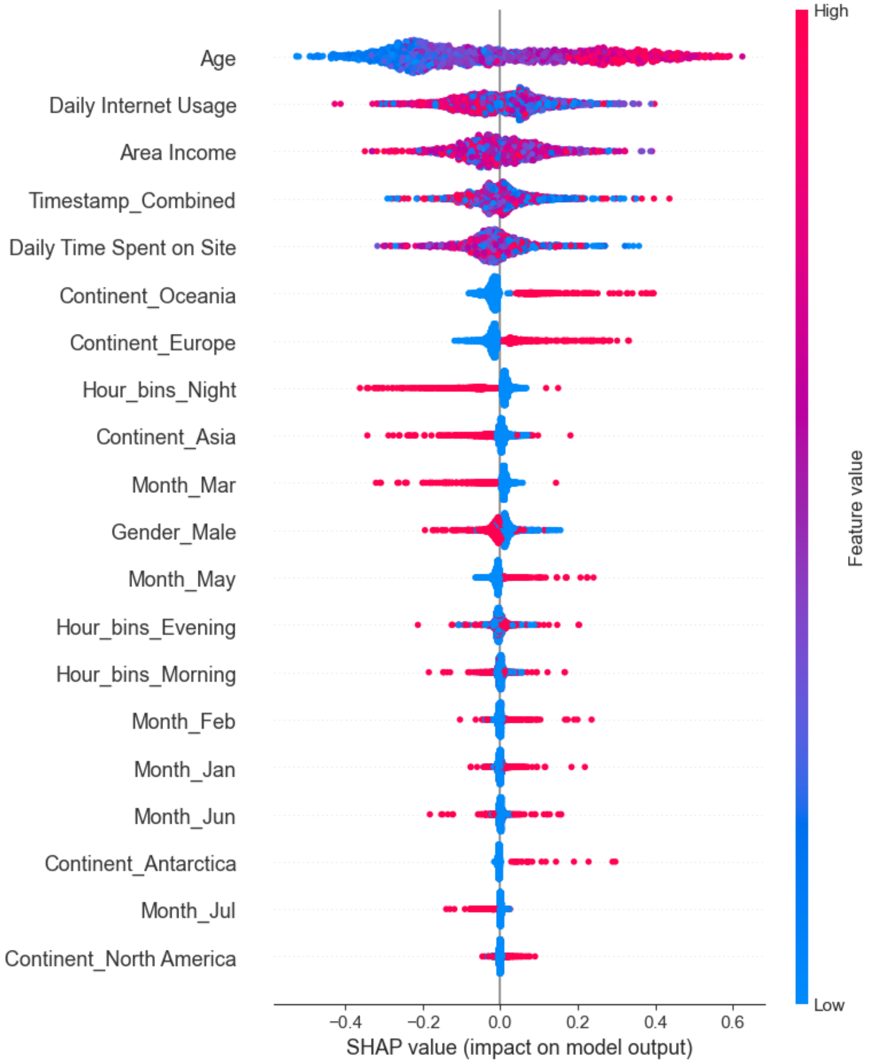
- n_estimators: 160
- max_depth: 8

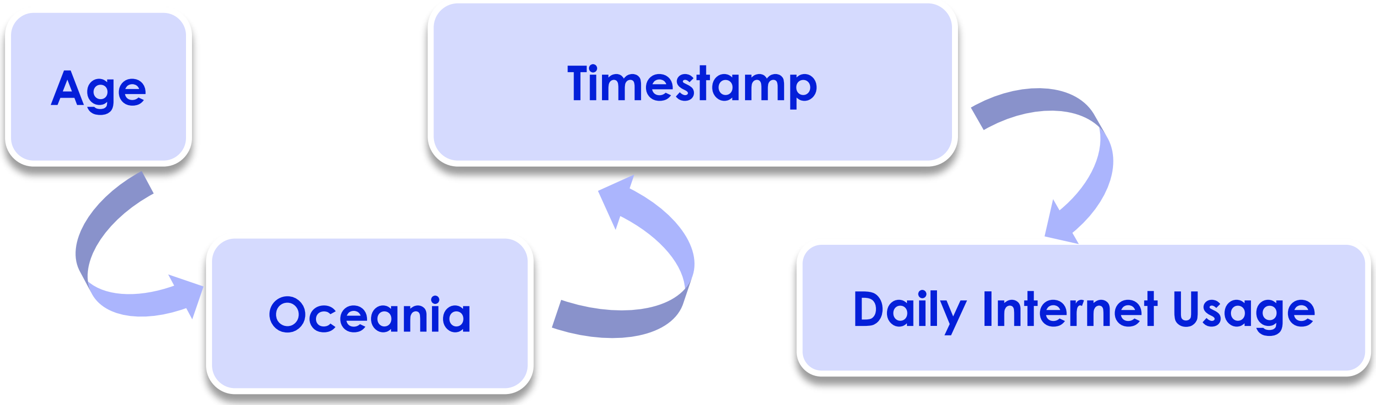




# TREE-BASED MODELS

## XGBoost



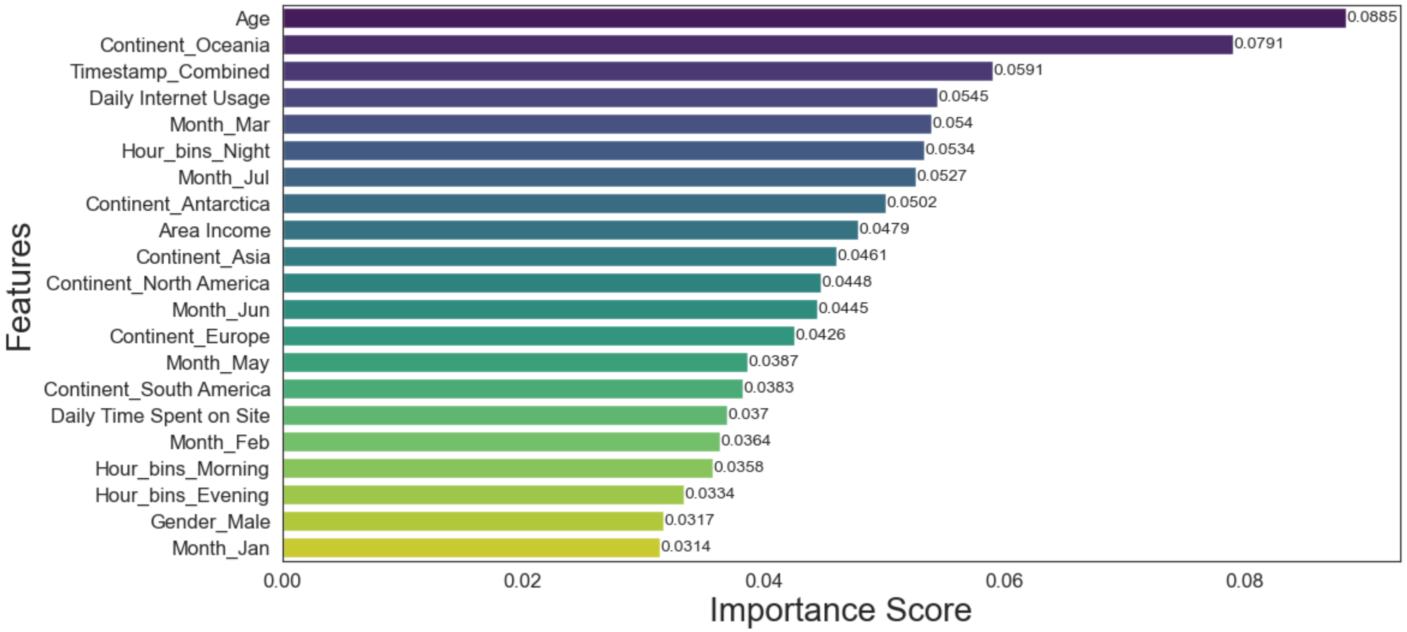


Age

Timestamp

Oceania

Daily Internet Usage



# MODEL EVALUATION

(UNSEEN DATA)

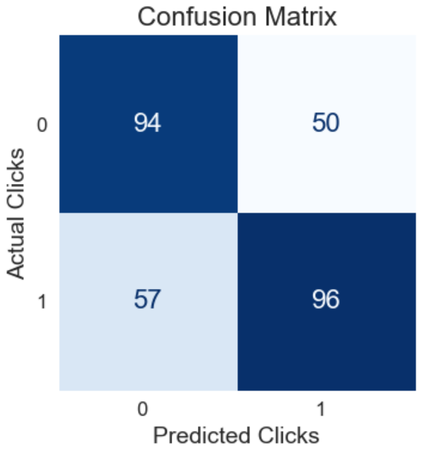
# MODEL EVALUATION

## Logistic Regression (Baseline)

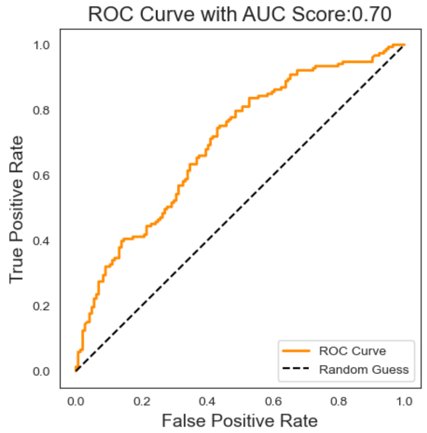
Accuracy: **0.63**

**ROC_AUC: 0.70**

F1 Score: **0.64**



Confusion Matrix



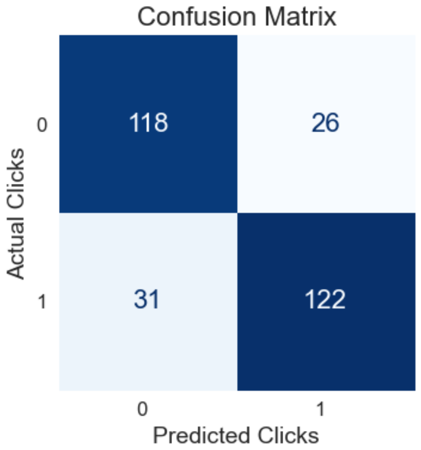
ROC Curve with AUC Score:0.70

## Random Forest (Tuned)

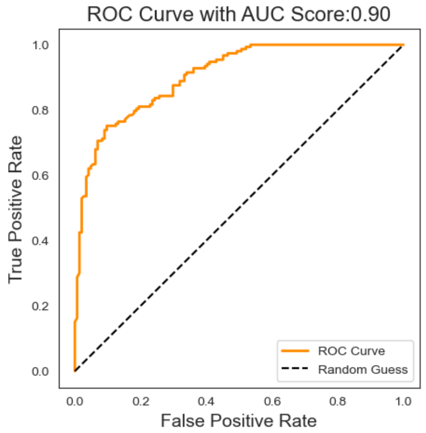
Accuracy: **0.80**

**ROC_AUC: 0.90**

F1 Score: **0.81**



Confusion Matrix



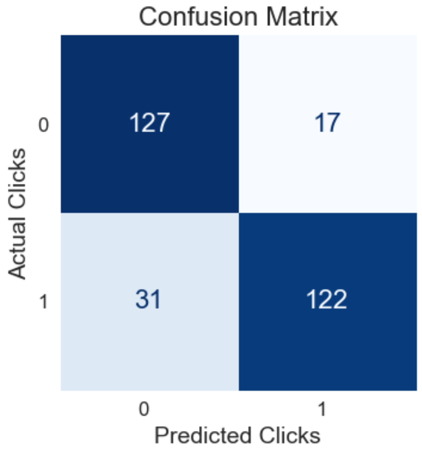
ROC Curve with AUC Score:0.90

## XGBoost (Tuned)

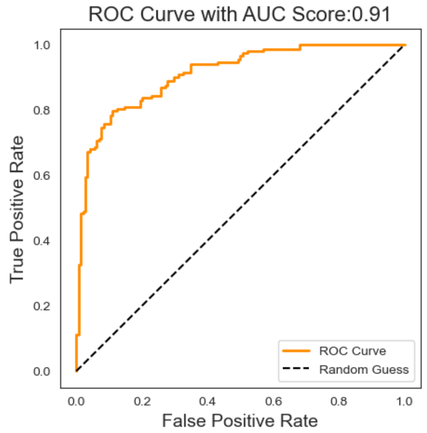
Accuracy: **0.83**

**ROC_AUC: 0.91**

F1 Score: **0.83**



Confusion Matrix



ROC Curve with AUC Score:0.91

# BUSINESS RECOMMENDATIONS

# MARKETING STRATEGIES

- **Target Older Audiences**
  Develop campaigns and messaging specifically targeted toward individuals in older age groups.

- **Focus on High Daily Internet Usage**
  Target these users through behavior-based segmentation and frequently visited platforms.

- **Optimize Ad Timing**
  Deliver ads during afternoon and evening hours when users are more likely to engage based on activity trends.

- **Prioritize High-Income Regions**
  Higher-income users show greater ad engagement. Target premium regions and promote high-value or luxury products.

- **Customize Campaigns by Regions**
  Analyze and tailor ad strategies for different regions (e.g., Oceania & Europe").

# THANK YOU

Li Wu

Instructor: Reza Moosavi