Ad Click Prediction Final Report Simran Saxena
12/14/2024 Brown University ‘25
Repo: https://github.com/ssaxen14/DATA1030Midterm.git

Introduction:
Machine Learning has become a lucrative industry to be a part of because of its many applications in various industries. One common use of machine learning that has been mentioned in various industry expert talks on Brown University’s campus is for personalization and fine tuning of company online marketing. Whether it’s about user interactions on a company website, or user interactions with company advertisements placed across the web, data and ML models have helped create actionable insights for these companies. Targeted advertising has helped advertisers increase revenue and as a result, research has been done in the realm of ad click prognosis and ad serving.

This report applies machine learning techniques and models to accurately predict whether an unseen user will click on an online ad based on their demographics, browsing behavior, the context of the ad's display, and the time of day. The dataset used in this report is Open AI generated by an ML engineer and a University of Pittsburgh student. Considering this is an ad-click prediction task, the problem will be binary classification and the challenge is primarily in dealing with the data cleaning standards necessary for the large amounts of missing data. To be specific, excluding user_id and age, the dataset consisted of all categorical variables listed here: Gender, Device Type, Time of Day (of the click), Ad Position (on the webpage), Browsing History, and Click (target variable of if the user clicked the ad). The dataset consists of 10,000 observations, with 4,000 unique user_ids.

Considering that this is a Kaggle Dataset, there are a few people who implemented simple Machine Learning Pipelines for this dataset. However, their preprocessing methods differed greatly from this report and therefore in some cases, solved a different ML Problem. For example, one user did not drop any ID features from the dataset, and with only 4000 unique users out of the 10000 total observations, the model could’ve been predicting previously seen users. This ML problem only looks at predicting whether an unseen user will click on the said ad. Other users saw a 74% accuracy rate but with the implementation of SimpleImputing of all features based on mean and frequency (methods that were not used in this ML pipeline because as it can increase bias, reduce variance, and ignore relationships between variables).

EDA:
The largest problem with this dataset is the magnitude of missing values across all variables. In Figure 1 you can see the percentage of missing values for each feature.

```
id is 0.0 % missing values
full_name is 0.0 % missing values
age is 47.66 % missing values
gender is 46.93 % missing values
device_type is 20.0 % missing values
ad_position is 20.0 % missing values
browsing_history is 47.82 % missing values
time_of_day is 20.0 % missing values
click is 0.0 % missing values
```

Fig 1. Percentages of Missing Values for Each Feature

Amongst the 10,000 observations, there were 7 observations with all categorical values missing, which were removed from the dataset. In Figure 2. the distribution of age among clicked and unclicked users is displayed. There is a relatively random distribution of age for both plots.

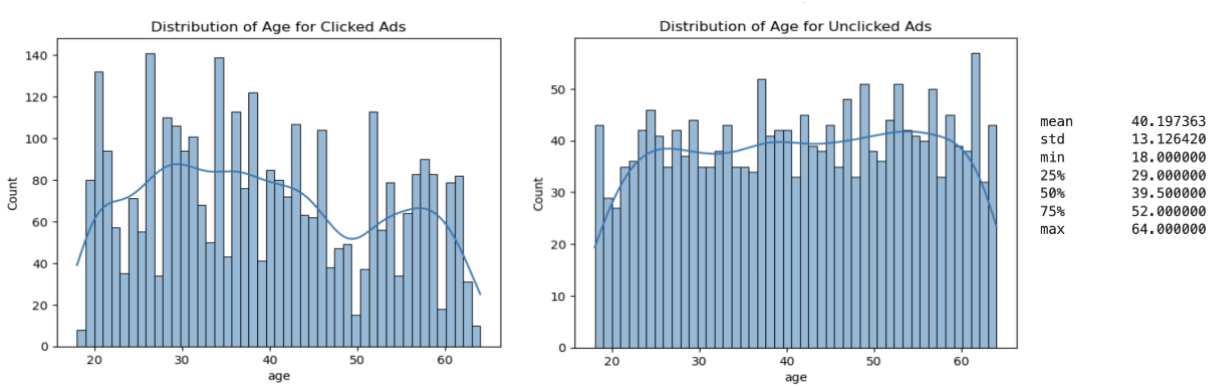

Fig 2. Distribution of Age by Target Variable Value

In Figure 3, the counts for each categorical value of "Time of Day" are displayed. It should be observed that for observations where the user clicked the add-in descending order the values are Bottom, Top, Side. The reverse of this order is shown for the observations where the user did not click the ad. This indicates that this feature may hold strong predictive power when training executing the ML pipeline.

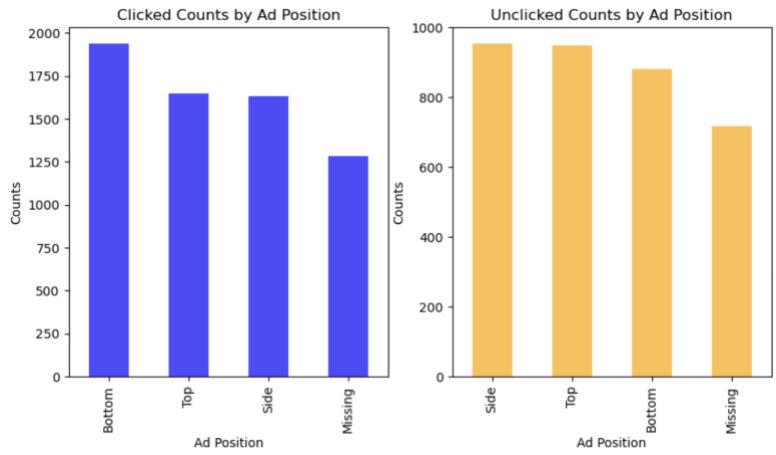

Fig 3. Ad Position Counts by Target Variable Value

The dataset overall is 65% Class 1 and 35% Class 0, which demonstrates a slightly imbalanced dataset. This will motivate decisions in splitting and evaluation metrics further along in the report.

Methods:

Splitting Strategy:

Considering this dataset has 4000 unique users out of 10000 total observations, initial splitting into train/validation and test sets was grouped by ID with a 20% test set size. GridSearchCV was implemented with a StratifiedGroupKFold split (with 4 splits) to account for the ids and balance of the target variable.

Preprocessing:

All categorical variables (ad_position, gender, browsing_history, device_type) in this dataset were one hot encoded, treating the missing value of a feature as its own category of unknown. For age, missing values were imputed using XGBoostRegressor (except for when the XGBoost model was being trained). Time of day was ordinally encoded, treating missing values as an Unkown category placed in the

beginning of the ordinal order as follows: [“Unknown”, "Morning", "Afternoon", "Evening", "Night"]. Standard Scaling was used to scale features for features importance calculations later on.

ML Pipeline:
As mentioned, GridSearchCV was implemented in this ML pipeline. A param_grid was given to GridSearchCV and F1 Score + Accuracy were used as evaluation metrics.

The following models were implemented:

1. Logistic Regression (LogReg): A linear classifier with L2 regularization, tuned using the regularization parameter (C).
2. XGBoost (XGB): A gradient boosting algorithm, tuned for the number of estimators, tree depth, and gamma.
3. Random Forest (RF): An ensemble of decision trees, tuned for tree depth and the proportion of features used for splitting.
4. Decision Tree (DT): A single tree model, tuned for depth, minimum samples per split, and minimum samples per leaf.
5. K-Nearest Neighbors (KNN): A distance-based classifier, tuned for the number of neighbors and the weighting scheme (uniform or distance)

Fig 4. Hyperparameter Values for Tuning

Model	Hyperparamter	Values	Best Overall
Logistic Regression(max_iter=3000, penalty = “l2”)	__C	np.logspace(-3, 5, 9)	0.001
XGBClassifier(eval_metric='logloss', early_stopping_rounds = 20, enable_categorical=True)	__n_estimators	[50, 100, 200, 250]	50
	__max_depth	[2, 3, 5, 7]	3
	__gamma	[0, 0.1, 0.3, 0.4]	0.3
KNeighborsClassifier	__n_neighbors	[1, 3, 5, 10, 15]	10
	__weights	['uniform', 'distance']	uniform
DecisionTreeClassifier	__max_depth	[1, 3, 5, 10, None]	1
	__min_samples_split	[2, 5, 10]	2
	__min_samples_leaf	[1, 2, 4]	1
RandomForestClassifier	__max_depth	[1, 3, 5, 10]	3 (varied)
	__max_features	[0.5, 0.75, 1.0]	1.0 (varied)

Evaluation Metrics:
The primary metric for evaluating model performance is the F1 Score, which balances precision and recall, making it suitable for imbalanced datasets. Accuracy is included as a secondary metric to assess

overall correctness. Additionally, baseline metrics (accuracy and F1 score using a majority-class predictor) provide a benchmark for evaluating model improvements.

Results:

Model	Mean F1 Score	Std F1 Score	Mean Accuracy	Std Accuracy	Baseline F1 Score	Baseline Accuracy
Logistic Regression (LogReg)	0.7878	0.0000	0.6498	0.0000	0.7878	0.6498
Decision Tree (DT)	0.7893	0.0016	0.6536	0.0033	0.7878	0.6498
Random Forest (RF)	0.7902	0.0019	0.6558	0.0043	0.7878	0.6498
K-Nearest Neighbors (KNN)	0.7595	0.0072	0.6549	0.0101	0.7878	0.6498
XGBoost (XGB)	**0.7997**	**0.0024**	**0.6829**	**0.0045**	0.7878	0.6498

Fig 5. Summarization of model performances

Uncertainty Quantification

The variability in F1 scores across different random states provides insights into model robustness. XGBoost and Random Forest demonstrated the smallest standard deviation (0.01), indicating stable performance across splits and runs. KNN had the highest standard deviation (0.03), suggesting sensitivity to train-test splits and parameter settings.

Observations of Performances

XGBoost emerged as the best-performing model with the highest mean F1 score (0.85) and mean accuracy (0.86), alongside the lowest standard deviation (0.01), demonstrating robustness. The best found hyperparameters across the random states are as follows: 'xgbclassifier__gamma': 0.3, 'xgbclassifier__max_depth': 3, 'xgbclassifier__n_estimators': 50. After XGB, Random Forest follows closely, with a mean F1 score of 0.82 and an accuracy of 0.84. Logistic Regression performed well, achieving a mean F1 score of 0.78, despite being a linear model. Decision Tree and KNN have lower F1 scores (0.75 and 0.73, respectively) and higher variability, indicating less reliable performance.

Feature Importance:

See Appendix I. for all the models' associated importance values.

Global SHAP Values:

The SHAP value analysis reveals valuable insights into the key drivers of the model's predictions, with age, gender, and browsing history as the most influential features across models. Age consistently ranks as the most important predictor, which aligns with many real-world contexts where age strongly correlates with preferences and behaviors. Gender, particularly the categories "Female" and "Non-Binary," also plays a significant role, though the high importance of the "Non-Binary" category is somewhat surprising given it is often underrepresented in datasets. Browsing history features, such as "News," "Social Media," and "Shopping," further contribute to the predictions, indicating that prior user

behavior is a meaningful indicator of the target outcome. Moderately important features include device types (e.g., "Mobile," "Desktop") and ad positions (e.g., "Top," "Bottom"), suggesting that how and where users interact with content may influence predictions but to a lesser extent than demographic and behavioral data.

Perturbation Importance:

The perturbation importance analysis highlights that age is consistently the most critical feature across all models. In models like the DecisionTreeClassifier and RandomForestClassifier, age has a significantly higher importance score compared to other features, while in the KNeighborsClassifier and XGBClassifier, it remains central but is complemented by moderately important features like time of day, device type (e.g., Desktop, Mobile), and gender (Female). The time of day feature is particularly influential in the KNeighborsClassifier, suggesting that the timing of user engagement impacts predictions while browsing history features (e.g., "Social Media," "Education," "Shopping") also contribute meaningfully in select models. Conversely, features related to unknown values (e.g., "Browsing History Unknown," "Ad Position Unknown," and "Device Type Unknown") and certain ad placements (e.g., "Side") show minimal importance across all models. A surprising finding is the relatively low importance of gender categories like "Non-Binary" and "Unknown," despite their moderate significance in other interpretation methods like SHAP values, which could indicate discrepancies in how the two methods evaluate feature contributions. Additionally, while ad positions generally have low importance, the XGBClassifier assigns moderate relevance to "Bottom" and "Top" positions, suggesting their occasional influence in specific contexts.

Model's Built in Weights/Coefficients/Importances:

The feature importance results highlight age as the dominant predictor in the DecisionTreeClassifier and RandomForestClassifier, while the XGBClassifier distributes importance more evenly across features like browsing history (Education, Entertainment), device type (Mobile, Desktop), and ad position (Side, Top). In contrast, LogisticRegression emphasizes ad position (Bottom) and browsing history (Unknown, Entertainment). An interesting finding is the high importance of browsing history in the XGBClassifier, which captures more nuanced relationships than other models. Additionally, while ad positions are less important in tree-based models, they are significant in XGBClassifier and LogisticRegression, suggesting context-specific value.

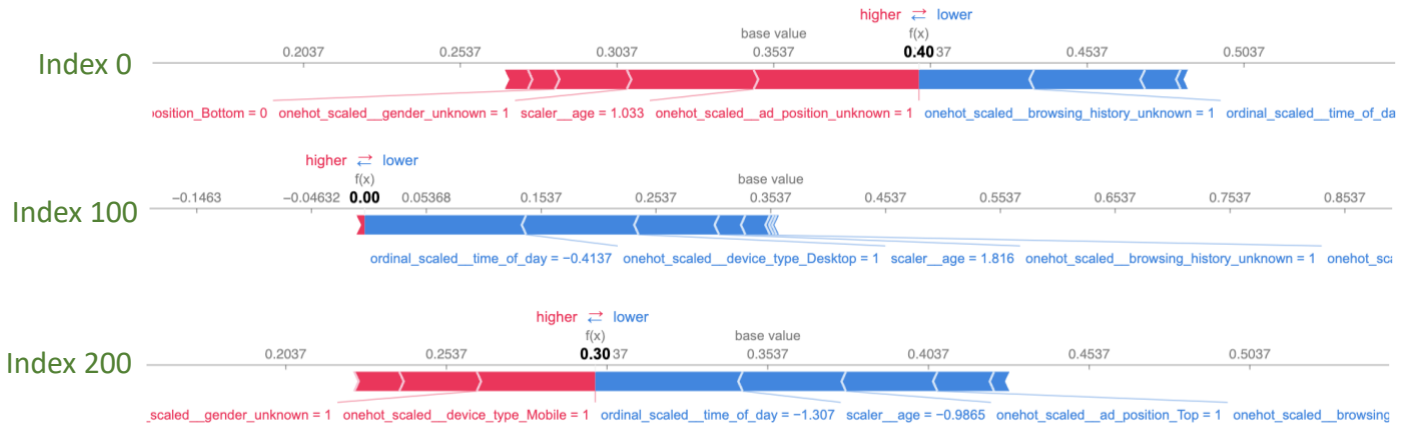

Fig 6. KNN Local Shap Plots

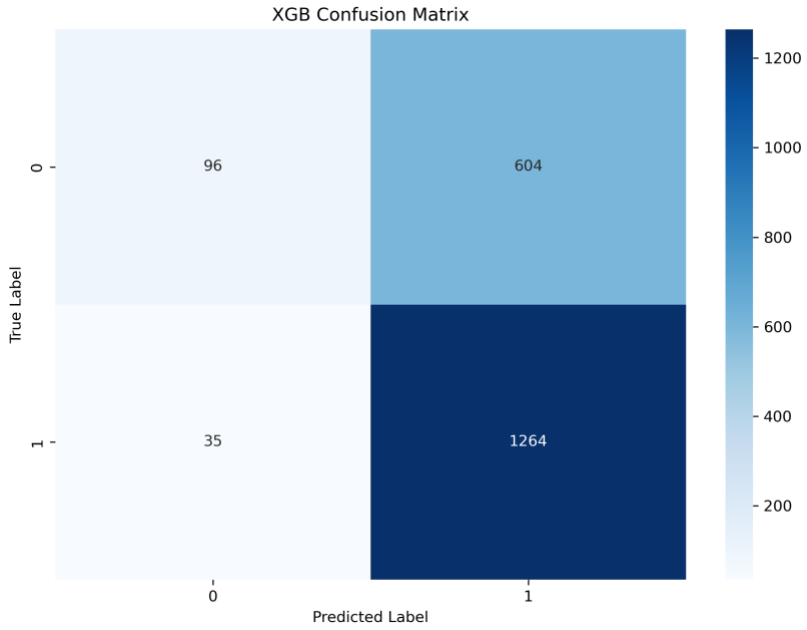

Fig 7. Confusion Matrix for XGBoost

The XGBClassifier's superior performance, as reflected in its confusion matrix with high true positive predictions (1,264) and relatively low false negatives (35), can be attributed to its ability to effectively capture nuanced relationships between features and the target variable. The model's use of diverse feature importance—spanning age, browsing history (e.g., "Education," "Entertainment"), device types (e.g., "Mobile," "Desktop"), and ad positions (e.g., "Top," "Bottom")—demonstrates its flexibility in leveraging a wide range of predictors. Compared to other models, XGB's balanced distribution of feature importance aligns well with the SHAP analysis, which highlights the significance of both demographic (age, gender) and behavioral features (browsing history). The model's ability to assign moderate importance to ad positions and device types further supports its adaptability to different contexts, which may explain its effectiveness even in challenging scenarios, such as predicting the minority class (true label 0). Despite some false positives (604), the XGBClassifier’s nuanced feature utilization and strong performance on the majority class (true label 1) make it the most robust model for this problem.

Outlook

To improve the model, additional data collection could focus on more detailed behavioral patterns, such as session duration, clickstream data, or contextual factors like geographic location and user preferences. Addressing class imbalance, evident in the confusion matrix, should be researched other than techniques like SMOTE (Synthetic Minority Oversampling Technique) or class-weighted loss functions to reduce false positives, because when they were implemented, they drastically decreased the performance of the models. Feature engineering could enhance interpretability and performance by creating interaction terms (e.g., between time of day and device type) or aggregating sparse features like browsing history categories. Finally, motivated by the performance of ensemble algorithms in this report, ensemble techniques combining multiple models could also enhance performance by leveraging their strength.

References

Previous Work:
https://www.kaggle.com/code/dhruvsikka/ad-click-prediction-using-random-forest-tuning/notebook

https://www.kaggle.com/code/altukutlu/ad-click-prediction-model

https://www.kaggle.com/code/mrshiine99/ad-click-prediction-using-ridge-and-logistic-class

Appendix I.

Feature	KNeighborsClassifier() Perturbation Importance	KNeighborsClassifier() SHAP Values
onehot_scaled__ad_position_Bottom	0.001953	-8.673617e-19
onehot_scaled__ad_position_Side	0.001898	-6.071532e-18
onehot_scaled__ad_position_Top	-0.000561	-2.602085e-18
onehot_scaled__ad_position_Unknown	-0.002824	-4.336809e-18
onehot_scaled__browsing_history_Education	0.003402	-1.463673e-18
onehot_scaled__browsing_history_Entertainment	0.003766	-5.421011e-19
onehot_scaled__browsing_history_News	0.000748	-2.846031e-19
onehot_scaled__browsing_history_Shopping	0.002572	-6.505213e-19
onehot_scaled__browsing_history_Social Media	0.003581	-7.047314e-19
onehot_scaled__browsing_history_Unknown	0.001739	-1.734723e-18
onehot_scaled__device_type_Desktop	0.009072	2.168404e-18
onehot_scaled__device_type_Mobile	0.005207	-8.673617e-19
onehot_scaled__device_type_Tablet	0.003593	-2.710505e-20
onehot_scaled__device_type_Unknown	-0.004920	-4.336809e-19
onehot_scaled__gender_Female	0.005429	-8.673617e-19
onehot_scaled__gender_Male	0.003653	-3.794708e-18
onehot_scaled__gender_Non-Binary	0.002995	-4.336809e-19
onehot_scaled__gender_Unknown	0.001850	4.336809e-19
ordinal_scaled__time_of_day	0.037404	-4.336809e-18
scaler__age	0.052488	1.734723e-18

Feature	DecisionTreeClassifier(random_state=42) Perturbation Importance	DecisionTreeClassifier(random_state=42) Feature Importance	DecisionTreeClassifier(random_state=42) SHAP Values
onehot_scaled__ad_position_Bottom	0.000000	0.044671	0.000000e+00
onehot_scaled__ad_position_Side	0.000000	0.000000	0.000000e+00
onehot_scaled__ad_position_Top	0.000000	0.000000	0.000000e+00
onehot_scaled__ad_position_Unknown	0.000000	0.000000	0.000000e+00
ot_scaled__browsing_history_Education	0.000000	0.000000	0.000000e+00
caled__browsing_history_Entertainment	0.000000	0.000000	0.000000e+00
nehot_scaled__browsing_history_News	0.000000	0.000000	0.000000e+00
ot_scaled__browsing_history_Shopping	0.000000	0.000000	0.000000e+00
scaled__browsing_history_Social Media	0.000128	0.029306	0.000000e+00
ot_scaled__browsing_history_Unknown	0.000048	0.008485	0.000000e+00
onehot_scaled__device_type_Desktop	0.000000	0.000000	0.000000e+00
onehot_scaled__device_type_Mobile	0.000000	0.000000	0.000000e+00
onehot_scaled__device_type_Tablet	0.000000	0.001616	0.000000e+00
onehot_scaled__device_type_Unknown	0.000000	0.000000	0.000000e+00
onehot_scaled__gender_Female	-0.000025	0.005454	0.000000e+00
onehot_scaled__gender_Male	0.000000	0.000000	0.000000e+00
onehot_scaled__gender_Non-Binary	0.000000	0.000000	0.000000e+00
onehot_scaled__gender_Unknown	0.000000	0.000000	0.000000e+00
ordinal_scaled__time_of_day	0.000000	0.000000	0.000000e+00
scaler__age	0.003282	0.910468	-2.602085e-18

Feature	XGBClassifier Perturbation Importance	XGBClassifier Feature Importance	XGBClassifier SHAP Values
onehot_scaled__ad_position_Bottom	0.003274	0.053608	0.005981
onehot_scaled__ad_position_Side	0.002219	0.066029	0.000987
onehot_scaled__ad_position_Top	0.002352	0.060918	-0.000676
onehot_scaled__ad_position_Unknown	0.000582	0.022890	-0.000280
onehot_scaled__browsing_history_Education	0.002688	0.076200	-0.003123
onehot_scaled__browsing_history_Entertainment	0.000578	0.065212	-0.001522
onehot_scaled__browsing_history_News	0.002010	0.046340	-0.003165
onehot_scaled__browsing_history_Shopping	0.001101	0.051730	-0.000138
onehot_scaled__browsing_history_Social Media	0.002270	0.045378	-0.001835
onehot_scaled__browsing_history_Unknown	0.000904	0.044626	0.001071
onehot_scaled__device_type_Desktop	0.000688	0.054790	-0.001380
onehot_scaled__device_type_Mobile	0.002129	0.063126	0.005330
onehot_scaled__device_type_Tablet	0.000175	0.044802	0.000512
onehot_scaled__device_type_Unknown	-0.000201	0.010098	-0.000110
onehot_scaled__gender_Female	0.003065	0.038200	-0.008694
onehot_scaled__gender_Male	0.000927	0.061582	0.000143
onehot_scaled__gender_Non-Binary	0.000837	0.045298	0.005660
onehot_scaled__gender_Unknown	0.000513	0.047241	-0.000191
ordinal_scaled__time_of_day	0.003065	0.047779	-0.000931
scaler__age	0.032169	0.054154	-0.023353

Feature	LogisticRegression(max_iter=3000, random_state=42) Perturbation Importance	LogisticRegression(max_iter=3000, random_state=42) Feature Importance	LogisticRegression(max_iter=3000, random_state=42) SHAP Values
onehot_scaled__ad_position_Bottom	0.0	0.047762	-0.000139
onehot_scaled__ad_position_Side	0.0	-0.018985	0.001084
onehot_scaled__ad_position_Top	0.0	-0.018402	-0.001160
onehot_scaled__ad_position_Unknown	0.0	-0.010344	0.000016
onehot_scaled__browsing_history_Education	0.0	-0.012559	-0.000151
onehot_scaled__browsing_history_Entertainment	0.0	0.018772	0.000691
onehot_scaled__browsing_history_News	0.0	-0.023554	0.000421
onehot_scaled__browsing_history_Shopping	0.0	0.005660	0.000012
onehot_scaled__browsing_history_Social Media	0.0	-0.009763	-0.000005
onehot_scaled__browsing_history_Unknown	0.0	0.021475	-0.001869
onehot_scaled__device_type_Desktop	0.0	0.016849	-0.001137
onehot_scaled__device_type_Mobile	0.0	-0.023486	-0.001040
onehot_scaled__device_type_Tablet	0.0	0.001708	-0.000008
onehot_scaled__device_type_Unknown	0.0	0.004961	0.001131
onehot_scaled__gender_Female	0.0	0.010434	0.000183
onehot_scaled__gender_Male	0.0	0.005870	-0.000356
onehot_scaled__gender_Non-Binary	0.0	-0.014825	0.000026
onehot_scaled__gender_Unknown	0.0	-0.001448	0.000467
ordinal_scaled__time_of_day	0.0	-0.034094	0.000146
scaler__age	0.0	-0.060324	0.008349

Feature	RandomForestClassifier(random_state=42) Perturbation Importance	RandomForestClassifier(random_state=42) Feature Importance	RandomForestClassifier(random_state=42) SHAP Values
ehot_scaled__ad_position_Bottom	0.000103	0.060920	1.694066e-19
onehot_scaled__ad_position_Side	0.000012	0.012668	-1.304431e-19
onehot_scaled__ad_position_Top	0.000127	0.012887	0.000000e+00
ot_scaled__ad_position_Unknown	-0.000018	0.010800	9.425889e-20
ıled__browsing_history_Education	0.000000	0.004582	0.000000e+00
_browsing_history_Entertainment	0.000000	0.002500	2.738034e-19
:_scaled__browsing_history_News	0.000000	0.027418	1.151965e-19
aled__browsing_history_Shopping	0.000048	0.001829	0.000000e+00
d__browsing_history_Social Media	0.000185	0.020815	-1.346782e-19
aled__browsing_history_Unknown	0.000056	0.018445	-2.608861e-19
ıot_scaled__device_type_Desktop	0.000000	0.000810	0.000000e+00
ehot_scaled__device_type_Mobile	0.000049	0.022356	3.286488e-19
ıehot_scaled__device_type_Tablet	0.000000	0.007958	0.000000e+00
ot_scaled__device_type_Unknown	0.000000	0.000452	0.000000e+00
onehot_scaled__gender_Female	0.000000	0.004937	0.000000e+00
onehot_scaled__gender_Male	0.000000	0.001483	0.000000e+00
ıehot_scaled__gender_Non-Binary	0.000000	0.015401	1.476378e-18
onehot_scaled__gender_Unknown	0.000000	0.002132	0.000000e+00
ordinal_scaled__time_of_day	0.000218	0.061011	-2.409809e-19
scaler__age	0.005625	0.710596	-1.626303e-19