October 2025

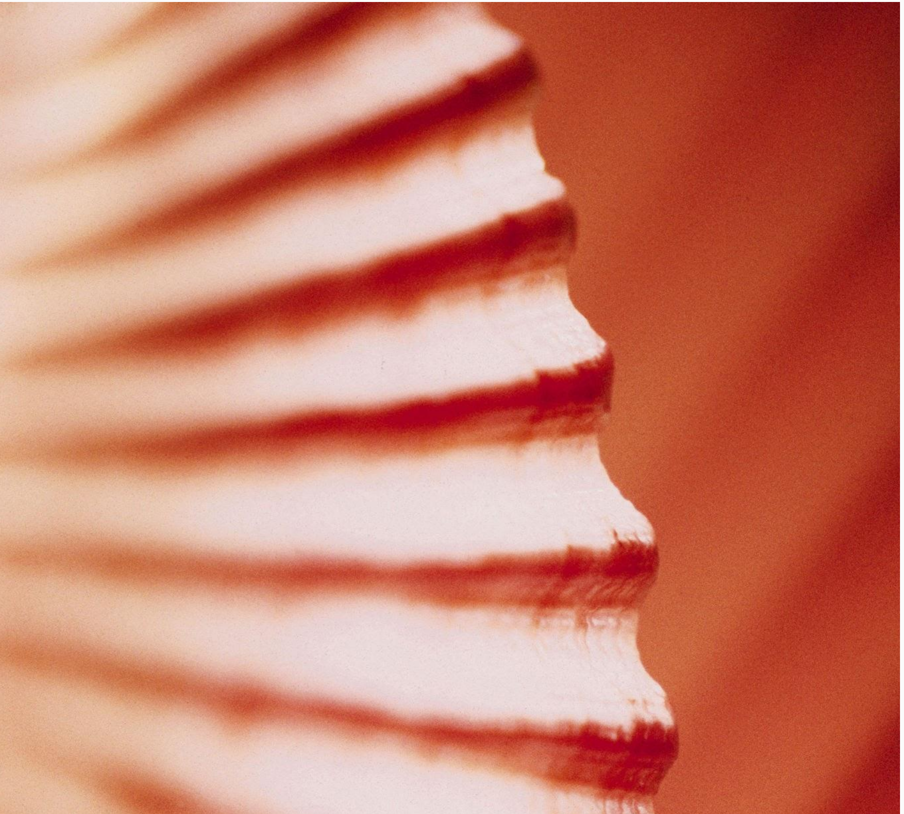
Analysis and Predictive Modeling *on ads and user data*

STATS 414: Generative AI

Setara Nusratty,
Nils Berzins,
Rohan Narasayya,
Lucy Lennemann

Agenda

October 2025



01 Problem Overview

02 Data Processing

03 Exploratory Data Analysis

04 Logistic Regression

05 Feature Selection with LASSO

06 XGBoost

07 Conclusions

01 Problem Overview

October 2025

Given information such as user gender, user engagement on news feeds, and ad placement ID, **can we predict whether an ad will be clicked**?

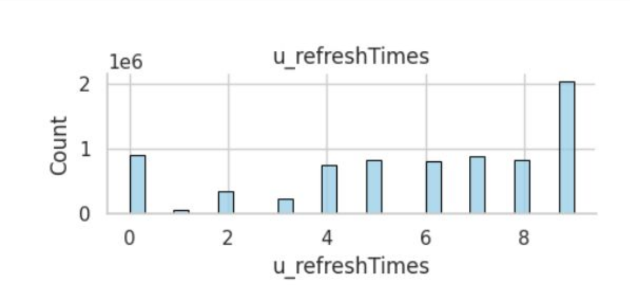
02 Data Processing

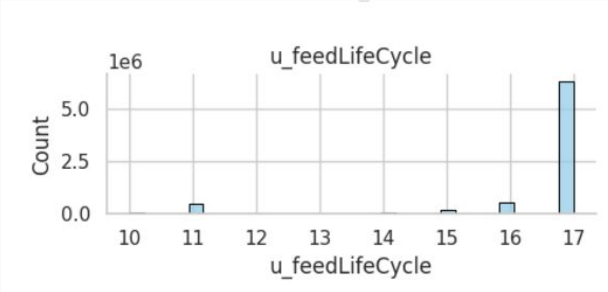
October 2025

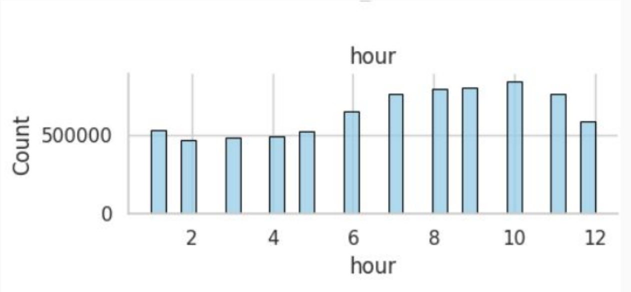
- Ads Dataset: 7 million rows, 35 predictors
- Change variable pt_d (time stamp) to date time format
- Standardize variables
- One hot encoded variables like age and gender as categorical
- Checked for missing values
- Engineered new features such as count of unique ads clicked and count of unique ads closed

03 Exploratory Data Analysis

October 2025

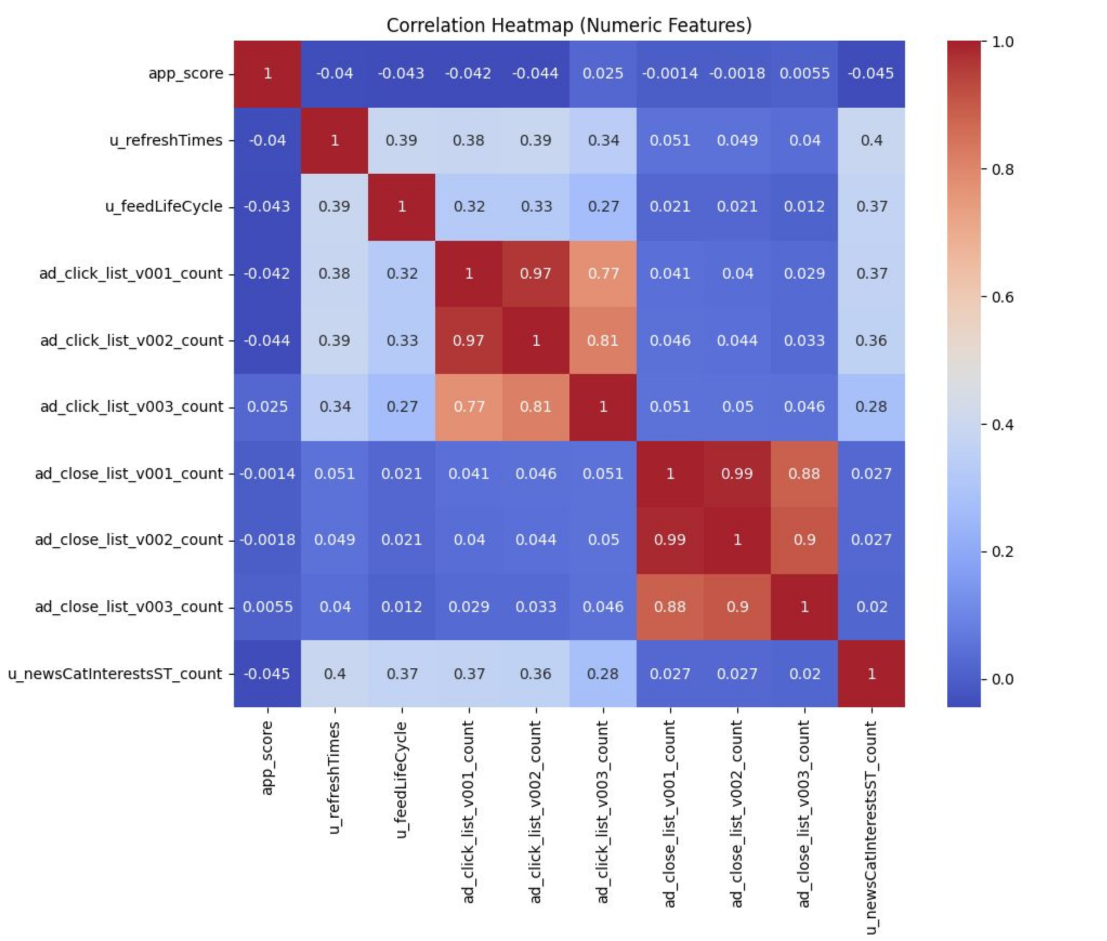






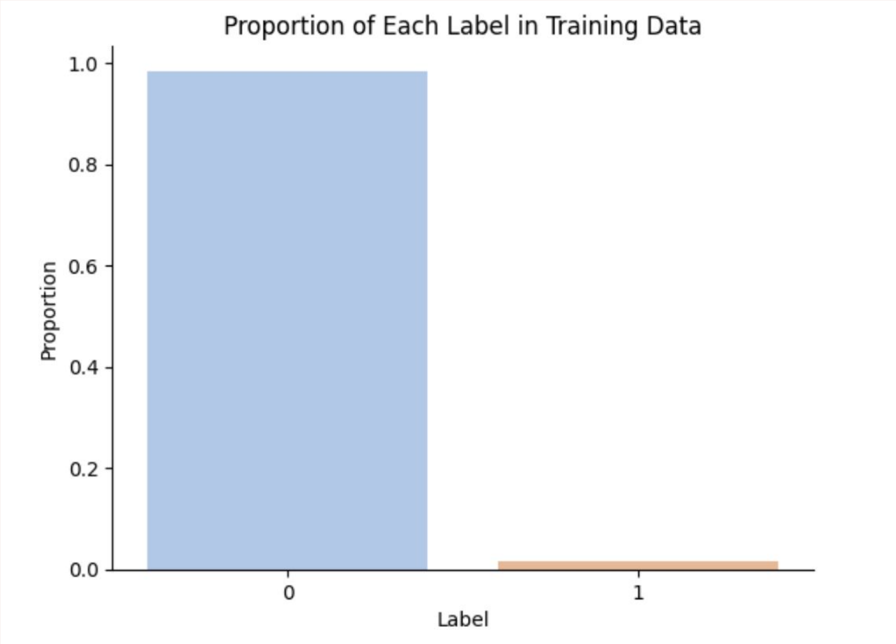
03 Exploratory Data Analysis

October 2025



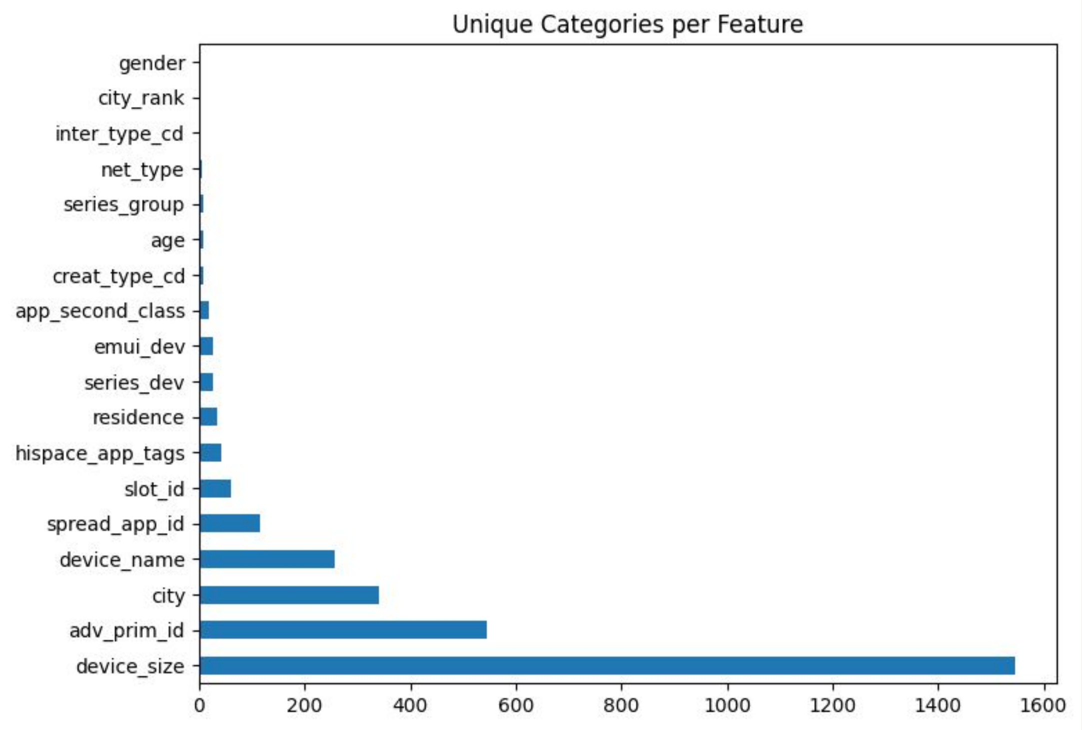
03 Exploratory Data Analysis

- Severely imbalanced dataset (1: ad clicked, 0: ad not clicked)
- Methods to deal with imbalance:
  - Undersampling from the majority class
  - Synthetic Minority Oversampling Technique for Nominal and Continuous (SMOTE-NC)
  - Adjusting model parameters



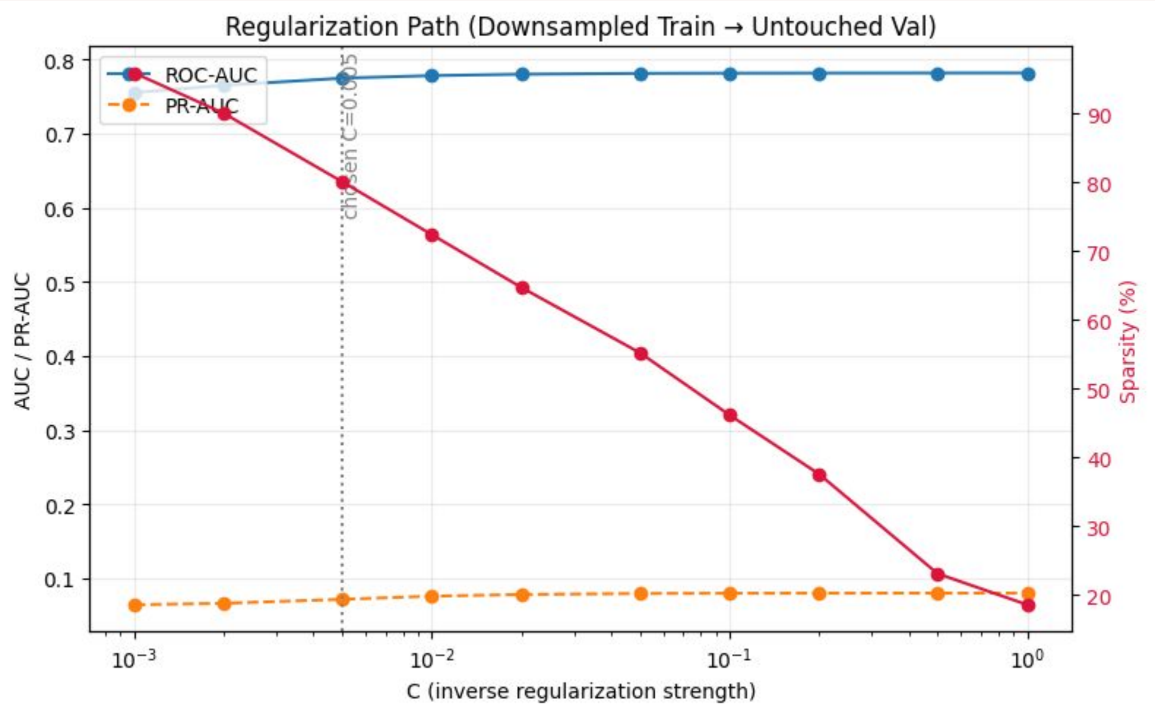
(5) Logistic Regression Pre-Processing

- Removed device_size, adv_prim_id, city, device_name
- Downsampled training data (6.1M to 200K) to balance classes



05 Feature Selection by LASSO

- C = 0.005
- 221 Coefficients Total
- 44 Nonzero after Lasso
- 80% Sparsity



(5) Evaluation Metrics

Across Threshold Metrics

- ROC-AUC: 0.77
- PR-AUC : 0.07

Ranks true clicks above non-clicks 77% of the time

Averages a precision of 7% across all recall levels

Confusion Matrix at Optimal Threshold of 0.80

	Pred 0	Pred 1
Actual 0	1473911	37366
Actual 1	18628	5199

- Precision: 0.12
- Recall: 0.22
- F1 Score: 0.16

(5) Logistic Regression Predictors

Top Positive Predictors

	feature	coef	odds_ratio
47	slot_id_46	1.017	2.77
59	slot_id_58	0.650	1.92
115	spread_app_id_197	0.610	1.84
66	slot_id_65	0.513	1.67
14	slot_id_13	0.458	1.58
29	slot_id_28	0.444	1.56
39	slot_id_38	0.385	1.47
36	slot_id_35	0.384	1.47
63	slot_id_62	0.327	1.39
193	app_second_class_18	0.246	1.28

Top Negative Predictors

	feature	coef	odds_ratio
217	net_type_4	-0.099	0.91
51	slot_id_50	-0.104	0.90
10	inter_type_cd_3	-0.109	0.90
60	slot_id_59	-0.132	0.88
31	slot_id_30	-0.186	0.83
0	u_feedLifeCycle	-0.234	0.79
1	u_refreshTimes	-0.530	0.59
214	creat_type_cd_10	-0.810	0.44
70	slot_id_69	-1.052	0.35
17	slot_id_16	-1.413	0.24

(5) Logistic Regression: Adding SMOTENC

- Experimented with an alternate way to deal with data imbalance
- SMOTENC to deal with both our numerical and categorical features
- Pipeline:
    - **Undersampling** so that majority to minority class has a ratio of 2:1
    - **SMOTENC** so that the end dataset has equal samples from ad clicked and ad not clicked
- Didn't result in significant model improvement

(5) Evaluation Metrics

Threshold Metrics: 0.5 & Optimal 0.824

- ROC-AUC: 0.7810 — Ranks true clicks above non-clicks 78% of the time
- PR-AUC : 0.0782 — Averages a precision of 7.8% across all recall levels

Confusion Matrix at Optimal Threshold of 0.824

	Pred 0	**Pred 1**
Actual 0	1471247	40030
Actual 1	18404	5423

- Precision: 0.119
- Recall: 0.228
- F1 Score: 0.1566

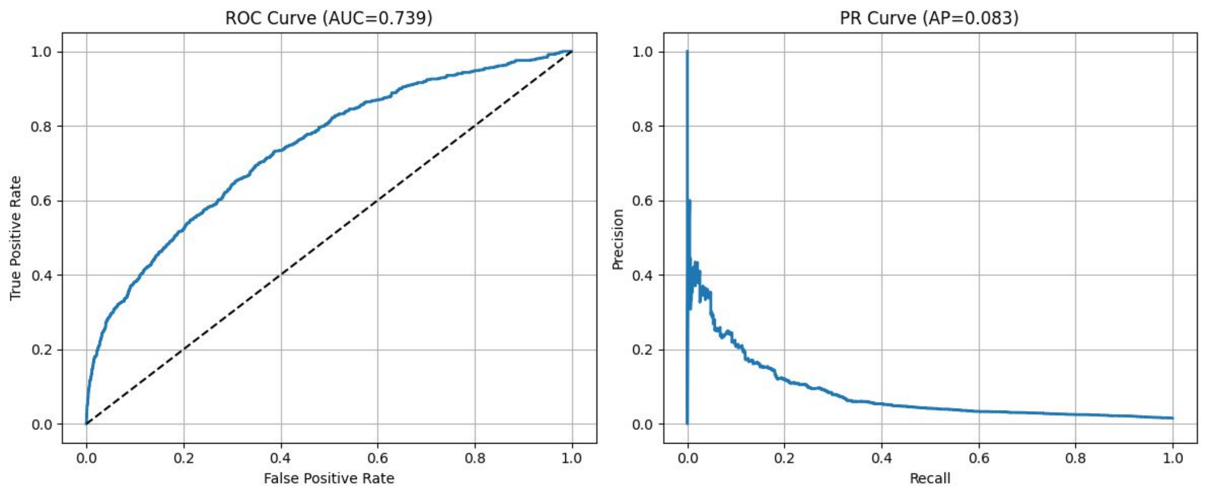
(6) XG Boost Pre-Processing

- Parsed **pt_d** as datetime and sorted by time
- List-like features converted to count features (e.g. ad_click_list_v001_count)
    - **Aggregated into**: click_count_mean, close_count_mean
- Standardized numeric features before applying SMOTENC
- Sampled 200,000 rows
    - Performed undersampling (5:1) on training set
    - 80/20 train/test split

Stage	**Total Rows**	**Class 0**	**Class 1**	**Ratio**
Original	200k	196,896	3,104	63:1
Train	160k	157,517	2,483	63:1
Test	40k	39,379	621	63:1
After Undersampling	14,898	12,415	2,483	5:1
After SMOTENC	24,830	12,415	12,415	1:1

(6) XG Boost with scale_pos_weight

- **scale_pos_weight**: 63.4
- **Model hyperparameters**:
  - n_estimators=400
  - max_depth=4
  - gamma=1.0
  - colsample_bytree=0.7
  - early_stopping=50
- **Had threshold** = 0.5
  - **Optimal threshold**: 0.988



(6) Evaluation Metrics

Threshold Metrics: 0.5 & Optimal 0.988

- ROC-AUC: 0.7388 — Ranks true clicks above non-clicks 74% of the time
- PR-AUC : 0.0830 — Averages a precision of 8% across all recall levels

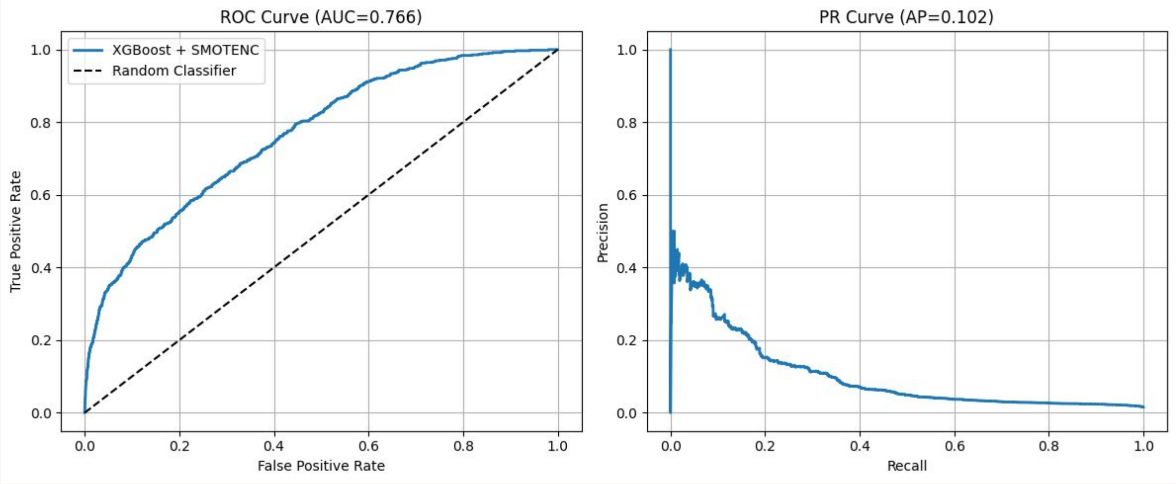
Confusion Matrix at Optimal Threshold of 0.988

	Pred 0	**Pred 1**
Actual 0	38738	641
Actual 1	511	110

- Precision: 0.1465
- Recall: 0.1771
- F1 Score: 0.1603

(6) XG Boost: Adding SMOTENC

- Applied SMOTENC to rebalance training to 1:1
- Removed scale_pos_weight
- **Model hyperparameters**:
  - n_estimators=400
  - max_depth=4
  - gamma=2.5
  - colsample_bytree=0.7
  - early_stopping=30
- **Had threshold** = 0.5
  - **Optimal threshold**: 0.826



(6) Evaluation Metrics

Threshold Metrics: 0.5 & Optimal 0.826

- ROC-AUC: 0.7659
- PR-AUC : 0.1021

Ranks true clicks above non-clicks 77% of the time

Averages a precision of 10% across all recall levels

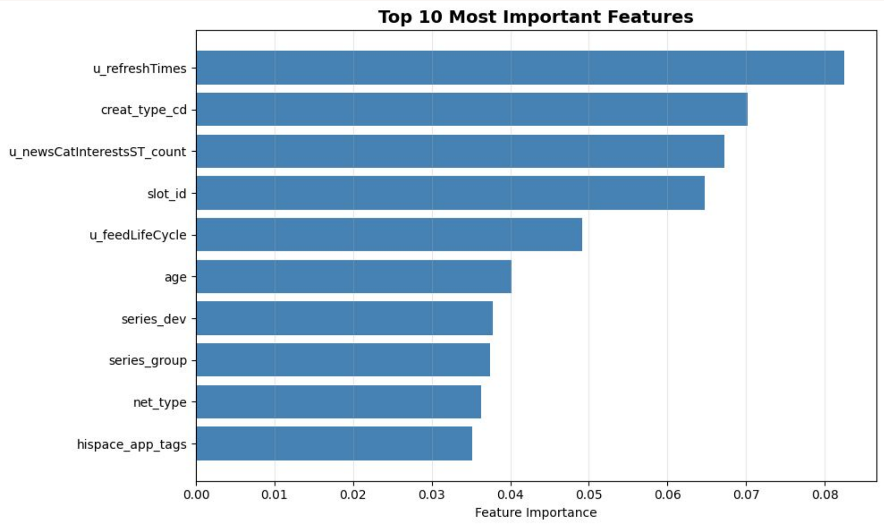
Confusion Matrix at Optimal Threshold of 0.826

	Pred 0	**Pred 1**
Actual 0	38932	447
Actual 1	511	110

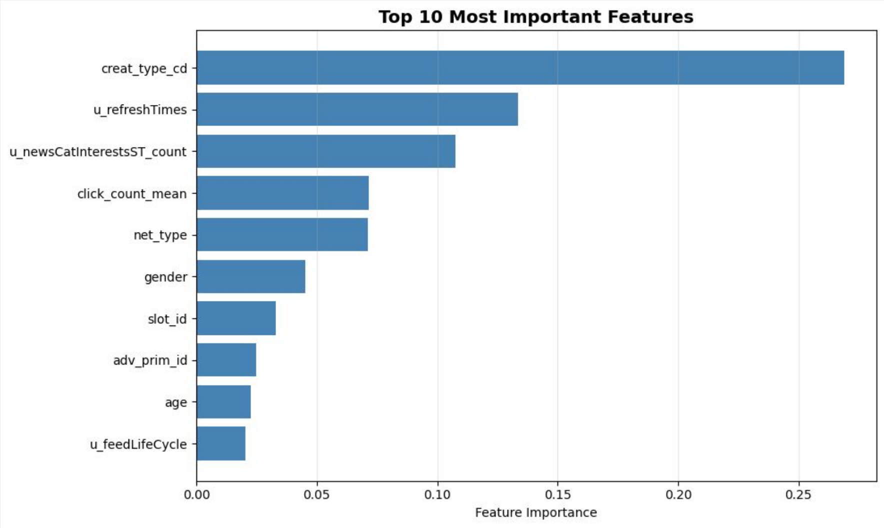
- Precision: 0.1975
- Recall: 0.1771
- F1 Score: 0.1868

(6) Feature Importance

Top Features: Simple XG Boost



Top Features: SMOTENC XG Boost



Engagement/activity, creative type, short term interest seem to be the top features in both models

Model Comparison

	F1 Score@0.5	ROC-AUC	PR-AUC
Random Classifier	0.03	0.5	0.02
Logistic Regression - undersampling	0.07	0.77	0.07
Logistic Regression - undersampling & SMOTENC	0.07	0.78	0.08
XGBoost - undersampling	0.05	0.74	0.08

(7) Conclusions

- Overall models perform better than the random classifier, which can reveal important insights for marketing companies about which factors get users to click
  - Ad placement and user engagement matter significantly whether a user clicks on an add
  - Content-type matching matters
- Limitations:
  - Imbalanced data leads to many false positives and low precision
  - Very marginal improvement in using SMOTENC
- Future work:
  - Fine-tuning hyperparameters
  - Further domain understanding about the variables and their meaning
  - Using the Feeds Dataset for a more feature-rich dataset
  - Utilizing cloud computing for model training and analysis

Thank you