December 2025

Predicting Ad Clickthrough-Rate (CTR)
on real and synthesized ads data

STATS 414: Generative AI

Setara Nusratty,
Nils Berzins,
Rohan Narasayya,
Lucy Lennemann

(1) Problem Overview

December 2025

What is the best approach to generate synthetic data to improve CTR classification?

(2) Data Pruning

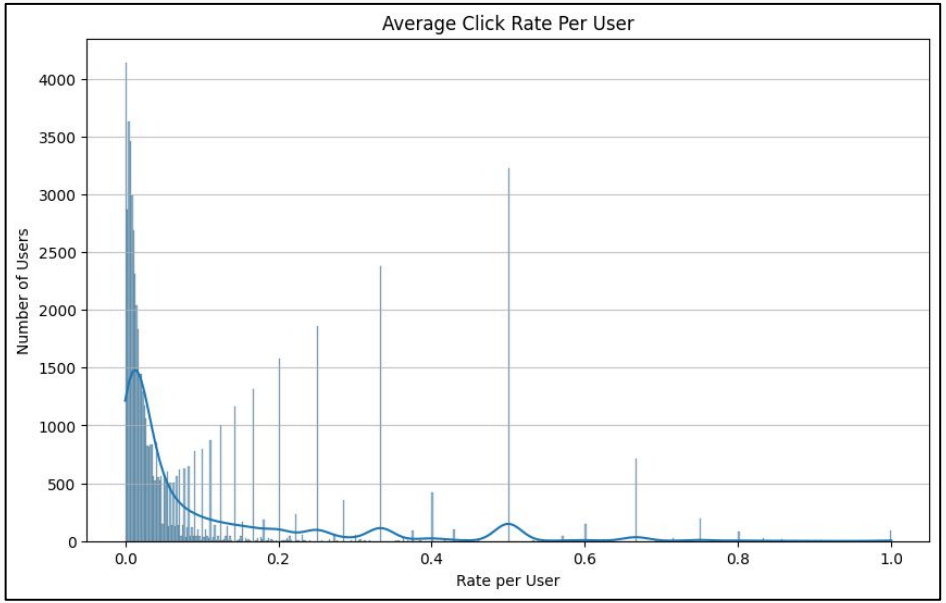
- **Goal: Reduce computational requirements by narrowing the scope of the problem**
- Original ads dataset: 7 million rows, 35 predictors
- Narrow down to top six task_ids, D1 to D6
- Focus on users who have at least clicked on an ad once

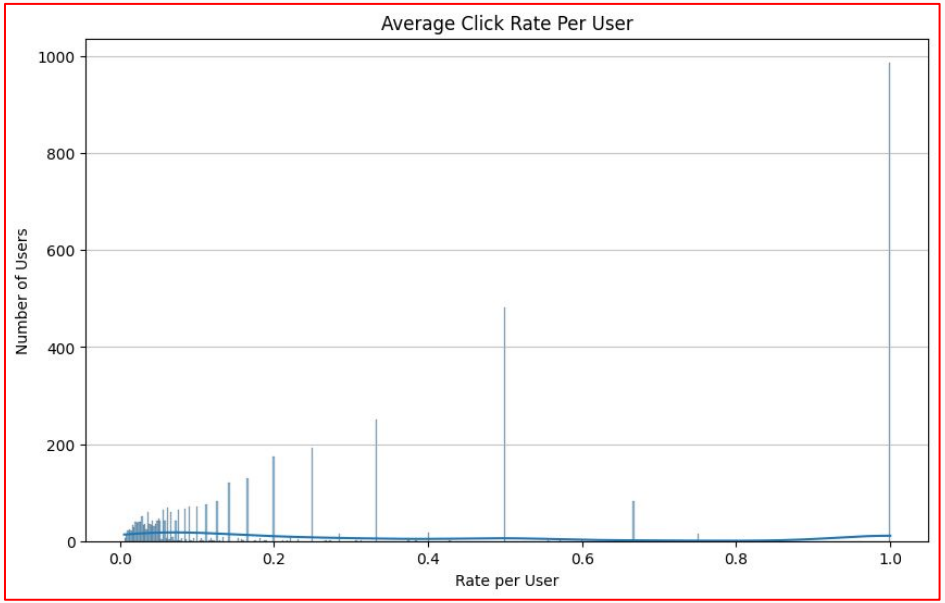
(2) Data Processing

- Converting variables to correct data type
  - pt_d → datetime
- Downcasted numerical data types, reducing memory footprint
- Engineer new count of interest category features
- Remove ID_fields for synthetic data generation
- Remove uninformative features through XGBoost importance analysis

(3) Data Analysis Pre/Post Pruning

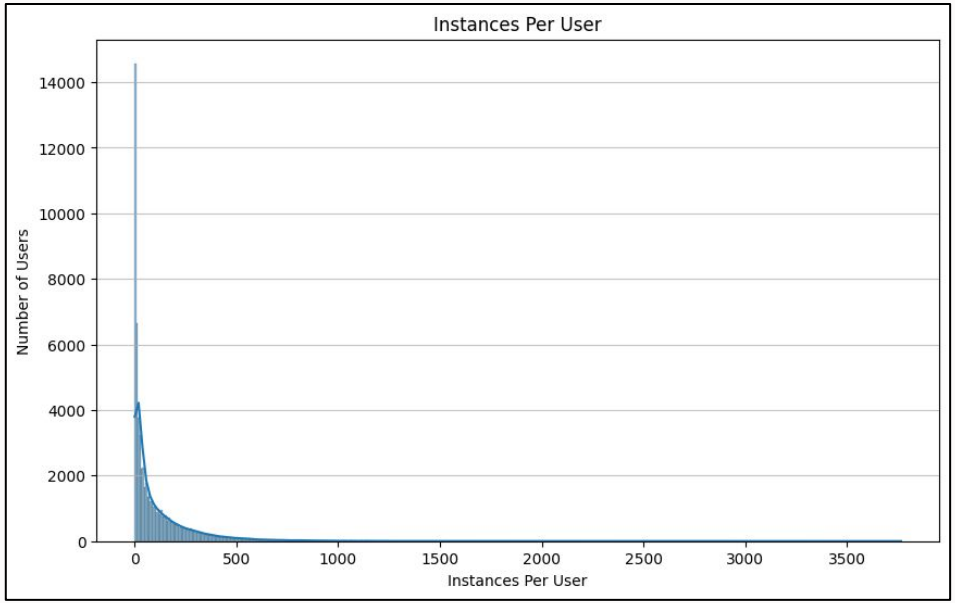
December 2025

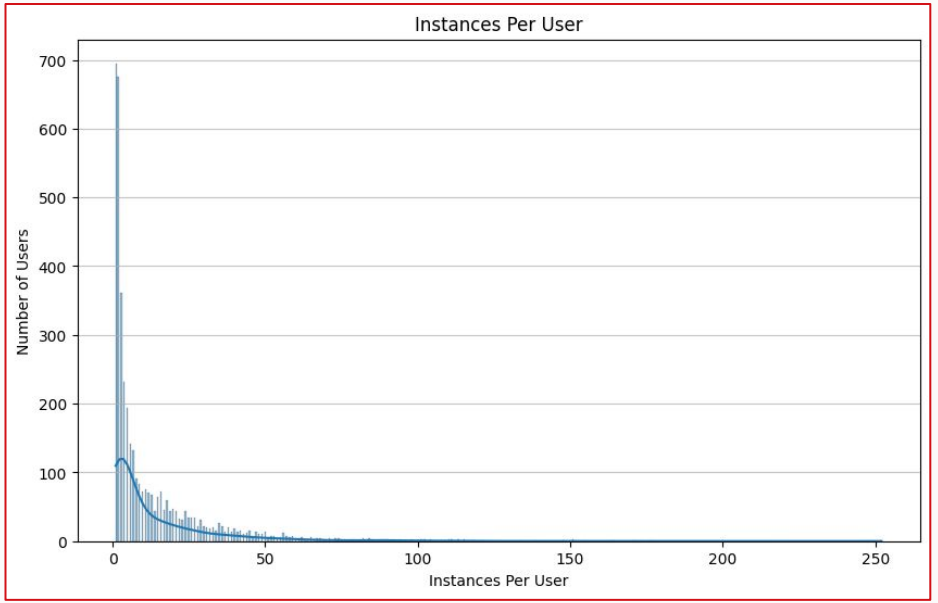




(3) Data Analysis Pre/Post Pruning

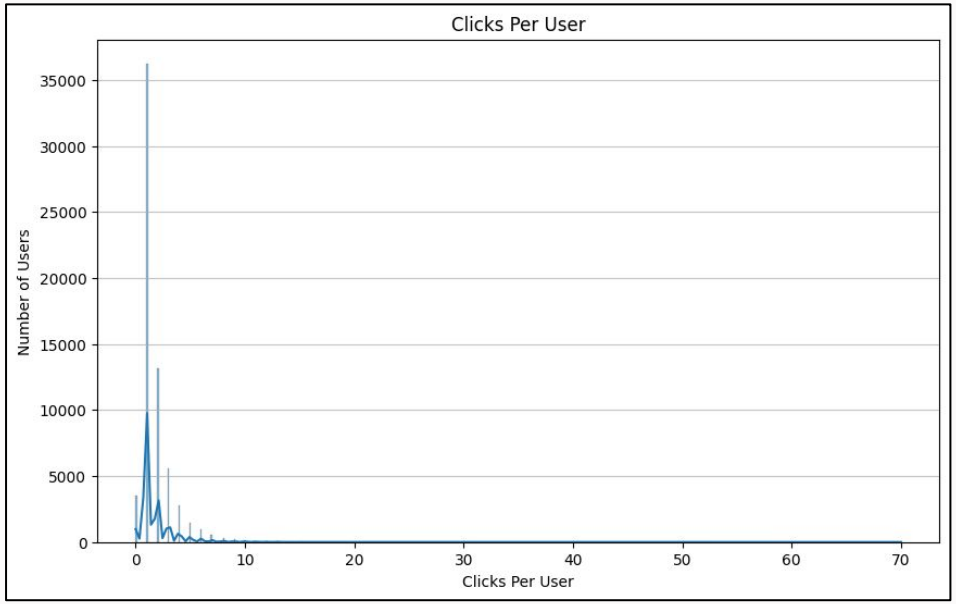
December 2025

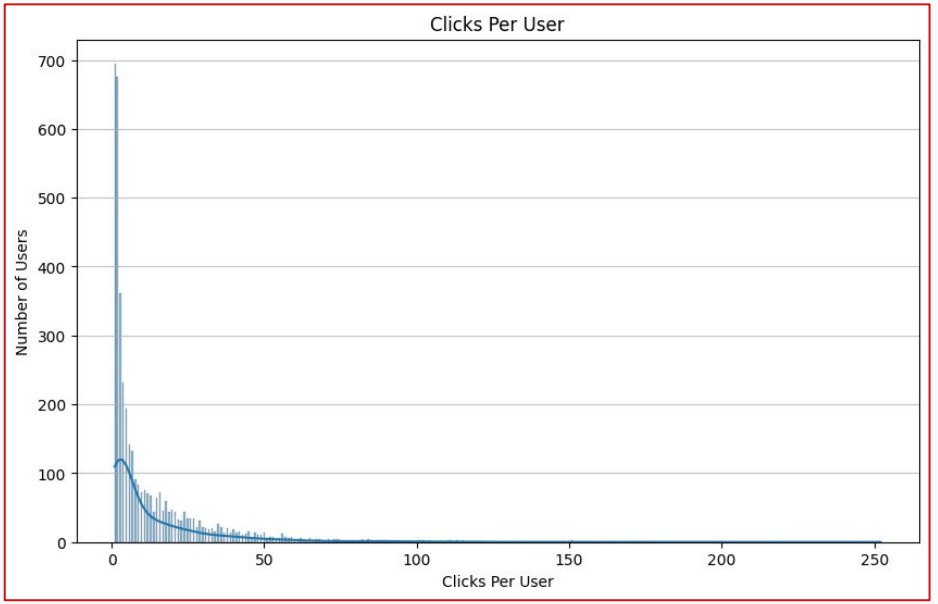




(3) Data Analysis Pre/Post Pruning

December 2025





(3.1) Data Analysis Post Pruning

December 2025

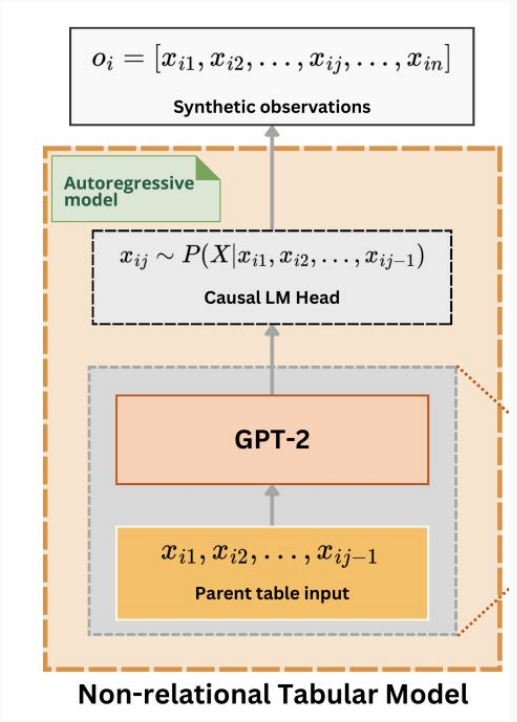
Dataset	task_id	rows_before	rows_after	ctr_before	ctr_after
1	22100	154812	4628	0.001815	0.060717
2	14584	126367	29359	0.014790	0.063660
3	34382	122094	2724	0.001589	0.071219
4	34975	92381	3880	0.001927	0.045876
5	31941	77470	12677	0.032167	0.196576
6	31996	69691	1555	0.002525	0.113183

Full Dataset CTR: 0.0947 (9.47%)

(4) Synthetic Data Generation: Realtabformer

December 2025

- Realistic Relational and Tabular Transformer
- Uses autoregressive transformer model, ChatGPT-2
- Changed hyperparameters to shorten training process
    - 75 epochs, turned off full sensitivity
- Train a different model for each subset of data D1 - D6 and generate a sample of 10,000 rows



(4) Synthetic Data Generation: CTGAN

- Conditional generative adversarial network
- Uses 2 neural networks, the generator and the critic, to improve the synthetic data
- Learns distributions more accurately due to techniques like conditional generation and oversampling of rare categories
- Train a different model for each subset of data D1 - D6 and generate a sample of 10,000 rows

(5) Problem Overview

How can we develop a framework to evaluate generated synthetic tabular data?

(6) Evaluation: Data Privacy

- Realtabformer already implements regularization to prevent data-copying
  - Uses target masking: artificially introducing missing values
  - Computes a stopping threshold to avoid overfitting using distance to closest record
- Further statistical tests to be evaluate data post-generation
  - DCR using Realtabformer method
  - Membership inference attacks (more comprehensive)

(6) Evaluation: Fidelity

- Conditional generative adversarial network
- Uses 2 neural networks, the generator and the critic, to improve the synthetic data
- Learns distributions more accurately due to techniques like conditional generation and oversampling of rare categories
- Train a different model for each subset of data D1 - D6 and generate a sample of 10,000 rows

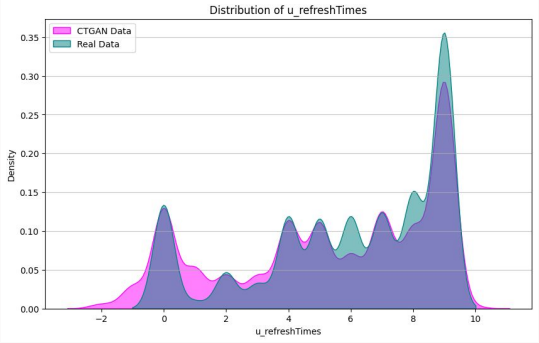
(6) Slide noting high-level synth data evaluation

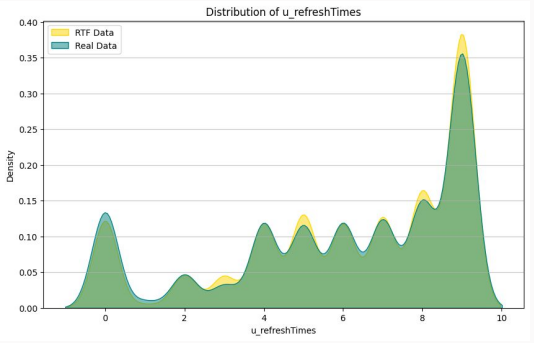
December 2025

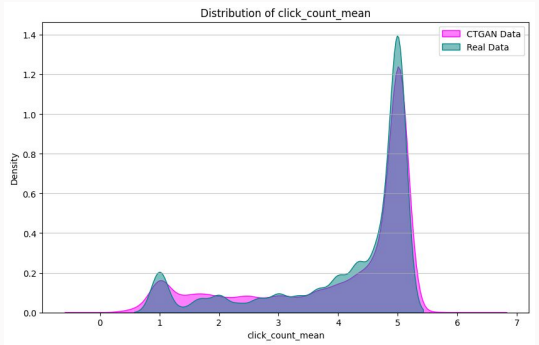
- Fidelity Evaluation:
  - Marginal dists can be estimated w/ total variational distance, correlations, and Cramer's V
  - Univariate Measures
    - mean, sd, skewness, etc
    - Chi-squared / total variation distance between empirical distributions
  - Pairwise
    - Numeric-Numeric -> matrix distances on between corr matrices
    - Categorical-Categorical -> Cramer's V
  - Classifier Modeling e.g. RandomForestClassifier modeling whether synthetic or not
    - If close to random guessing, P and Q are similar
  - PRD Precision-Recall on CTGAN and RTF separately
    - Drop to lower dimension space (PCA)
  - How close is the data to the real data?

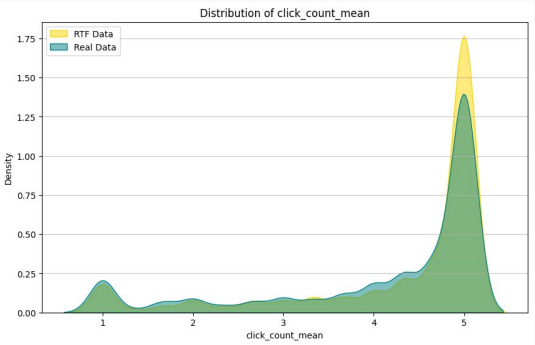
(6) Fidelity: Graphical Analysis

December 2025









(6) Fidelity: Total Variational Distance and Marginal Cramer's V

Numerical Features: Total Variational Distance

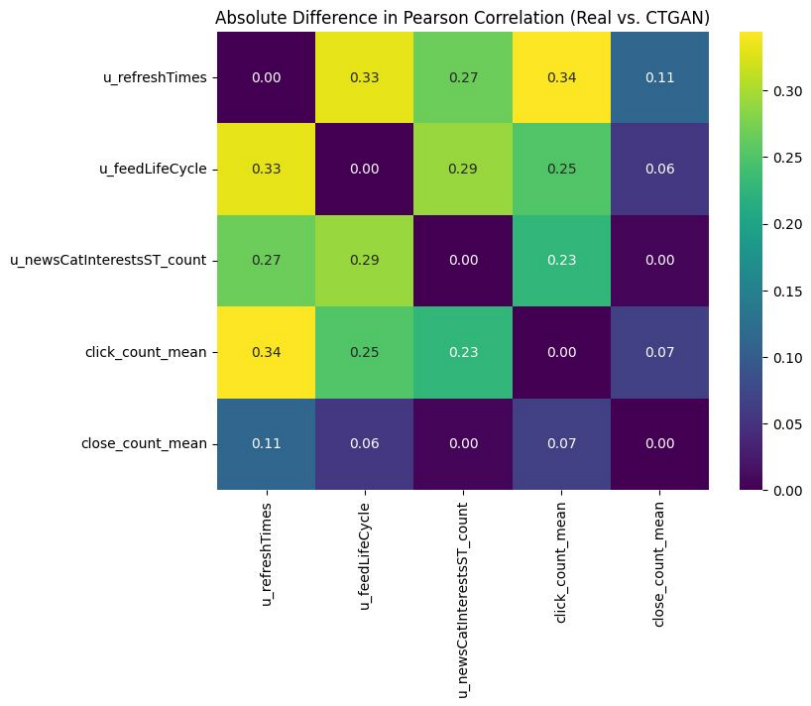
Feature	TVD_CTGAN	TVD_RTF
u_refreshTimes	0.101774	0.032038
u_feedLifeCycle	0.098143	0.014617
u_newsCatIntesrestST_count	0.093361	0.014055
click_count_mean	0.377356	0.087038
close_count_mean	0.037952	0.008097

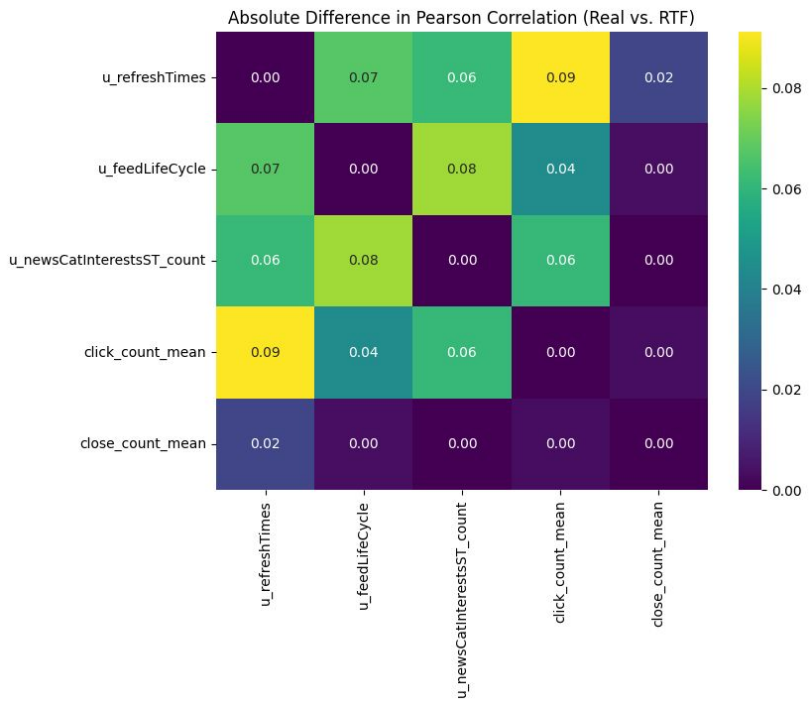
Categorical Features: Cramer's V

Feature	CramersV CTGAN	CramersV RTF
age	0.101708	0.094325
gender	0.026075	0.034176
city_rank	0.083767	0.086042
device_size	0.350771	0.368568
slot_id	0.464447	0.448289

(6) Fidelity: Absolute Correlation Differences

December 2025





(6) Fidelity: Random Forest Classifier

Detection of Synthetic Data Classification Report

Label	Precision	Recall	F1-Score	Support
0	0.78	0.72	0.75	16447
1	0.88	0.91	0.89	36000

Accuracy Score: 0.85067

Slide noting high-level synth data evaluation

December 2025

- Utility Evaluation:
  - Choosing downstream task (already doing, predicting classification 1/0)
    - Test different levels of synthetic data inclusion
    - ALSO, train on synthetic, test on real
  - Model Selection → Taking many models and seeing how introduction of synth data changes optimal model
    - Rank by performance
    - If performance is consistent across different synth data injection levels, **can actually use** synth data for true predictive modeling
  - Utility-interpretability trade off
  - How useful is the data for a task?

(7) Utility: Modeling Pipeline Overview

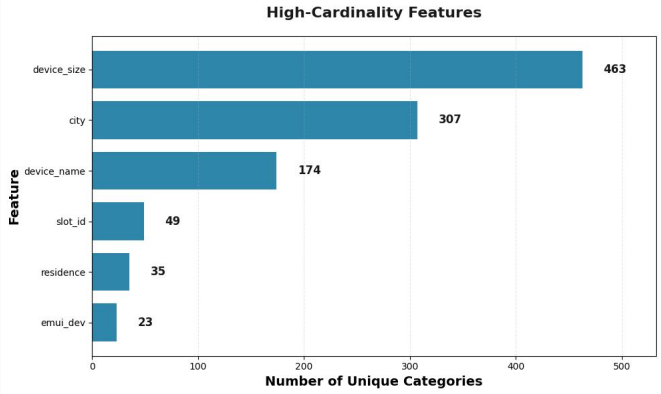
December 2025

- **Goal:** Compare utility of CTGAN vs RealTabFormer synthetic data
- **Two identical pipelines:** Same preprocessing, models, ratios
- **Data:** 54k real (60% train, 20% val, 20% test) + 60k synthetic (D1-D6 per generator)
- **Models:** L1 Logistic Regression and XGBoost
- **Experiments:** Real-only, Synthetic-only, Mixed (0.10×–1.64×)
- **Evaluation:** AUC and PR-AUC (test set)

(7) Logistic Regression Preprocessing

December 2025

- Numerical features → StandardScaler
- Low-cardinality features → One-Hot Encoding
- High-cardinality features → Target Encoding:
    - Minimum 20 unique values per feature
    - Smoothing (Pulls value toward global mean)
    - Clipping to ±3σ (Prevents any encoded value from going outside a realistic range)



(7) Logistic Regression Results

December 2025

Realtabformer

	Synthetic Ratio	Test AUC	PR-AUC
0	0.00× (Real Only)	0.77840	0.40977
1	0.10×	0.77868	0.4093
2	0.36×	0.77883	0.41354 ◀ best
3	0.61×	0.77838	0.40862
4	0.87×	0.77930	0.41112
5	1.13×	0.77887	0.40649
6	1.38×	0.77803	0.40701
7	1.64×	0.77815	0.40349
8	Synthetic Only	0.77033	0.39437

CTGAN

	Synthetic Ratio	Test AUC	PR-AUC
0	0.00× (Real Only)	0.77840	0.40977 ◀ best
1	0.10×	0.77955	0.40593
2	0.36×	0.77793	0.39458
3	0.61×	0.77701	0.38814
4	0.87×	0.77441	0.37847
5	1.13×	0.77190	0.37312
6	1.38×	0.77136	0.37609
7	1.64×	0.77111	0.37339
8	Synthetic Only	0.72978	0.29227

(7) Logistic Regression Interpretation

- City dominates model signal
- Device type & size matter
- Slot ID strongly predictive
- Other features have small effects

Rank	nice_name	coefficient	Odds Ratio
1	city	4.66150	105.794
2	device_size	2.08987	8.084
3	device_name	2.00163	7.401
4	slot_id	1.69733	5.459
5	type=cd_8	0.58251	1.791
6	u_refreshTimes	-0.40471	0.667
7	age_9	0.34999	1.419
8	u_newsCatInterestsST_count	-0.27054	0.763
9	type=6	0.22072	1.247
10	group=3	0.18961	1.209

(7) XGBoost Results

Realtabformer

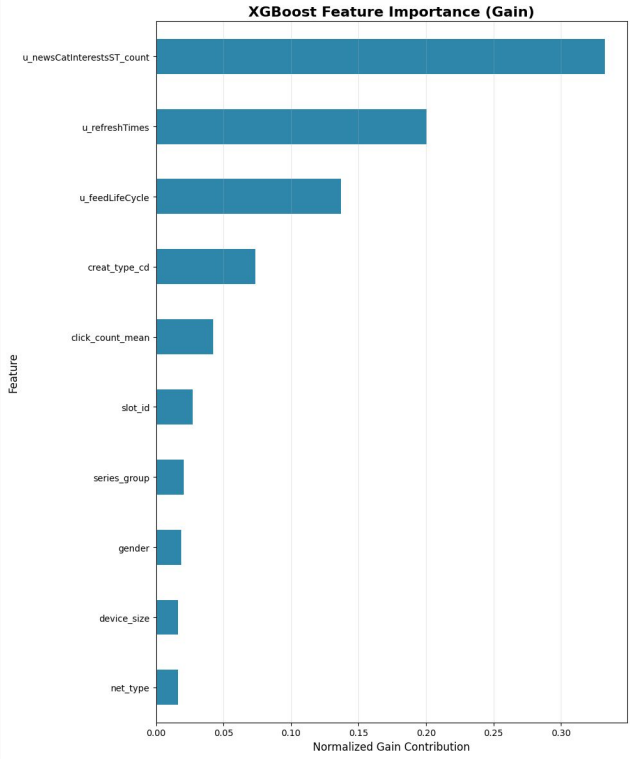
	Synthetic Ratio	Test AUC	PR-AUC
0	0.00× (Real Only)	0.78092	0.43334
1	0.10×	0.78306	0.43656
2	0.36×	0.78519	0.43858
3	0.61×	0.78573	0.44387
4	0.87×	0.78884	0.44473
5	1.13×	0.79093	0.44294
6	1.38×	0.79288	0.46273 ◀ best
7	1.64×	0.79101	0.44661
8	Synthetic Only	0.78976	0.44515

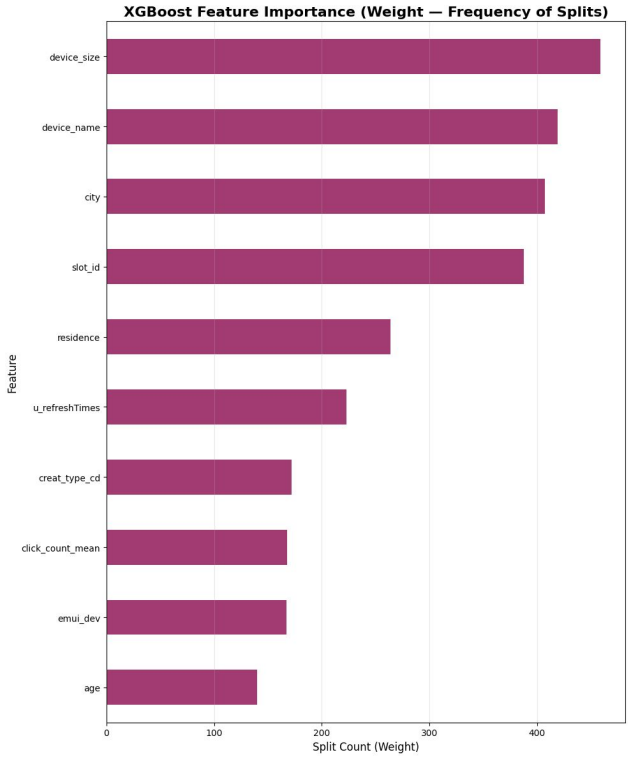
CTGAN

	Synthetic Ratio	Test AUC	PR-AUC
0	0.00× (Real Only)	0.78092	0.43334
1	0.10×	0.78380	0.43327
2	0.36×	0.78386	0.43234
3	0.61×	0.78367	0.43414
4	0.87×	0.78270	0.43602 ◀ best
5	1.13×	0.78095	0.43285
6	1.38×	0.78160	0.43246
7	1.64×	0.77896	0.42638
8	Synthetic Only	0.72610	0.31395

(7) XGBoost Feature Importance

December 2025



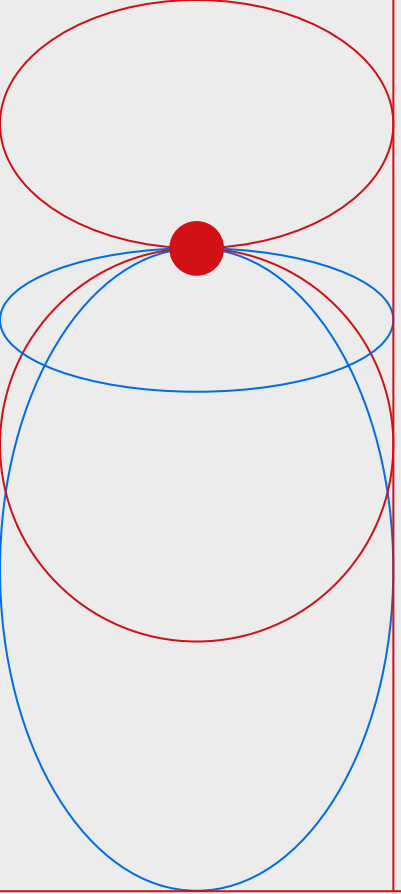


(8) Conclusions

1. Pruning improved CTR signal and reduced computational burden
2. The ideal synthetic-to-real ratios varied by model type
    - Logistic Regression performed best at very low synthetic ratios (0.0×–0.36×) and degraded steadily as synthetic proportion increased.
    - XGBoost performed well even with high synthetic ratios (0.87×–1.38×), showing that more synthetic data can improve a strong nonlinear model
    - REaLTabFormer tolerated higher ratios than CTGAN, which degraded faster
3. Synthetic data marginally improves model's prediction power over random classifier
4. **REaLTabFormer performed best across fidelity and utility**

(8) Limitations/Future Steps

- **Limitations:**
  - Training cost and instability — ReaLTabFormer required long training times and hyperparameters were adjusted to speed up the process
  - Evaluation limited to two synthesizers (CTGAN + RTF)
- **Future work:**
  - Run more comprehensive privacy evaluations
  - Experiment with more recent LLMs
  - Scale synthetic training to full dataset instead of subsets



Thank you