Conversion Attribution & Uplift Modelling:

Dataset:

https://ailab.criteo.com/criteo-uplift-prediction-dataset/

Aim:

Develop a machine learning or statistical modeling approach to dissect and attribute the contribution of each variable to the final purchase decision.

Approach:

To dissect and attribute the contribution of each variable in the final purchase decision (i.e., the "conversion"), we designed a multi-step approach combining both predictive modeling and interpretability techniques. Our goal was not only to build an effective classification model for conversions but also to uncover **which features most strongly influence the likelihood of a user converting**.

The dataset exhibited extreme class imbalance (very few positive conversions). To tackle this:

- We applied under sampling of the majority class (non-conversions) to rebalance the training set.
- We also experimented with focal loss and "scale_pos_weight" for models like XGBoost to emphasize harder-to-classify samples**.**
- XGBoost (with and without imbalance-aware tuning)
- **Logistic Regression**
- Gradient Boosting Classifier

Feature Contribution Analysis (Interpretability)

- We used **SHAP** (SHapley Additive exPlanations) for tree-based models.
- SHAP summary plots and bar charts revealed the global feature importance.
- For local, instance-level explanations, we used **LIME** on sampled test data.

Metric:

We prioritized recall slightly more than precision since our goal is to capture as many true converters as possible, even at the cost of a few false positives. This aligns with typical marketing or ad-tech use cases where maximizing conversion opportunities is more valuable than strictly minimizing misclassification. Used Metrics:

- F1
- Precision
- Recall

Uplift Modelling

To better understand and quantify the incremental impact of a treatment (e.g., exposure to a campaign) on user conversions, we implemented uplift modeling. Unlike traditional classification, which only predicts outcomes, uplift modeling estimates the causal effect of the treatment on individual users by comparing outcomes between treated and untreated groups.

Approach

We trained two types of uplift models:

- **Solo Model**: A single logistic regression classifier that models conversion probability by excluding the treatment indicator at training and evaluating uplift post-prediction.

- **Two-Model Approach**: Separate logistic regression models were trained for the treatment and control groups. Uplift was calculated as the difference in predicted probabilities between the treated and control groups.

Evaluation Metric: Cumulative Gain

We evaluated uplift performance using the **Cumulative Gain Curve**, which measures the **incremental gain in conversions** by targeting users ranked by their predicted uplift score. The gain is calculated as the difference in conversion rates between treated and untreated users cumulatively down the sorted list.