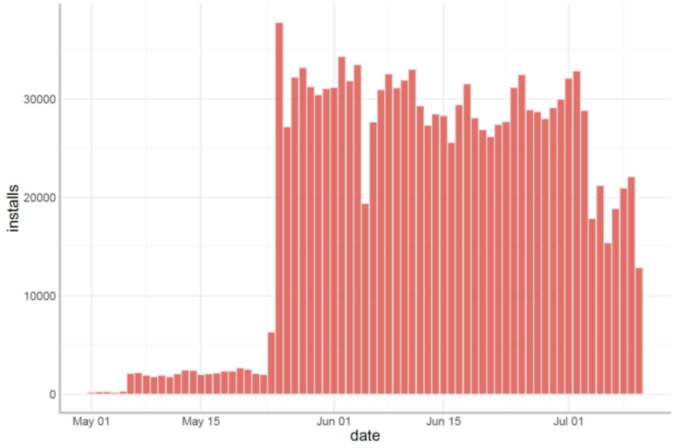
BUBBLEYE Coding Test

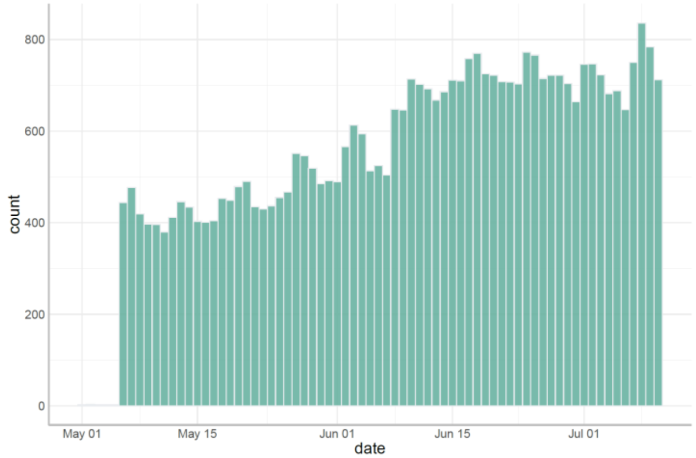
Chi-Chieh Huang

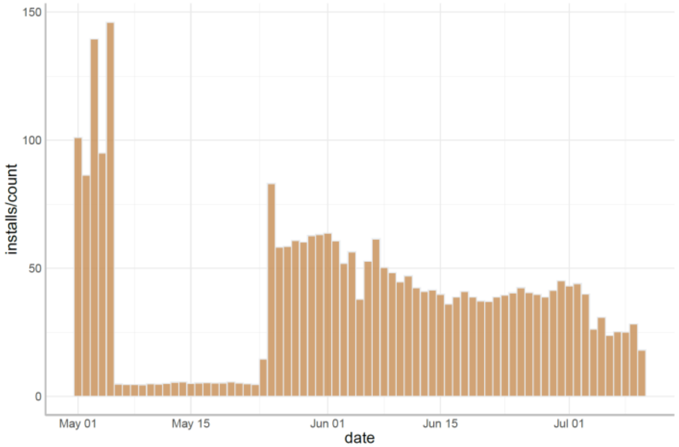
Exploratory Data Analysis

Descriptive Statistics

- Campaign: 50 unique
- Publisher: 288 unique
  - Campaign + Publisher: 3,419 unique
- Date：
  - Start: 2023-05-01
  - End: 2023-07-10

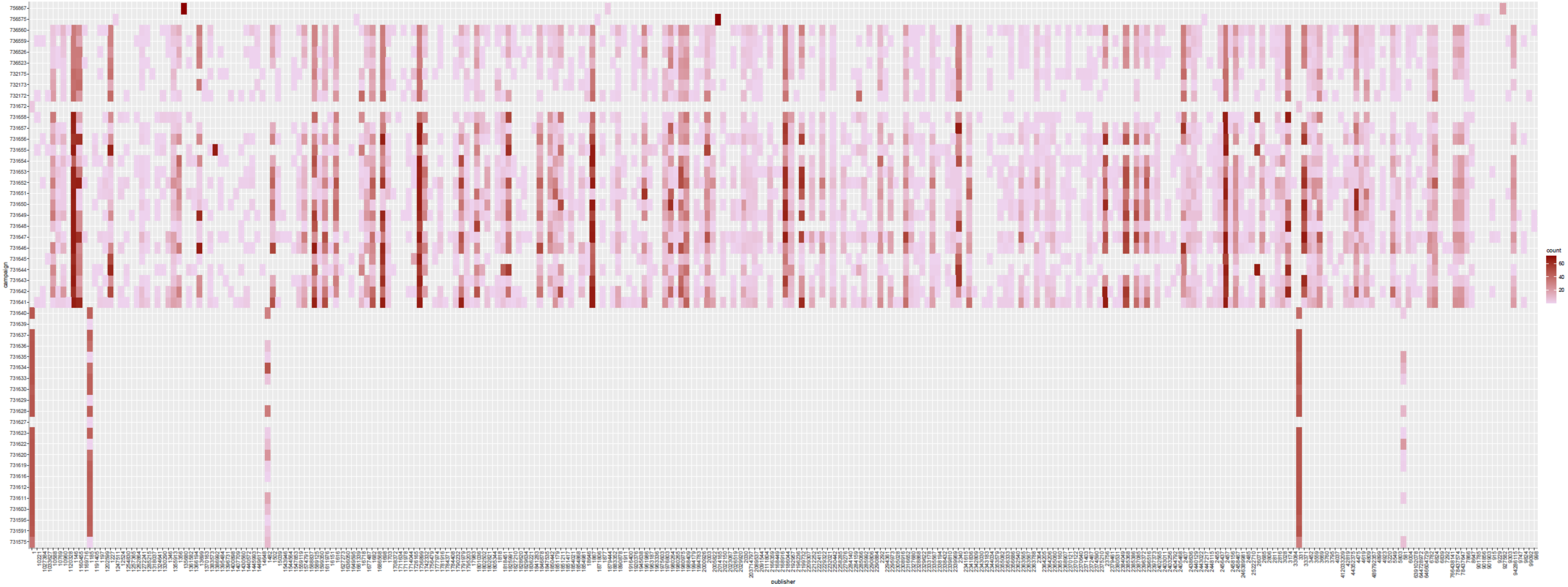






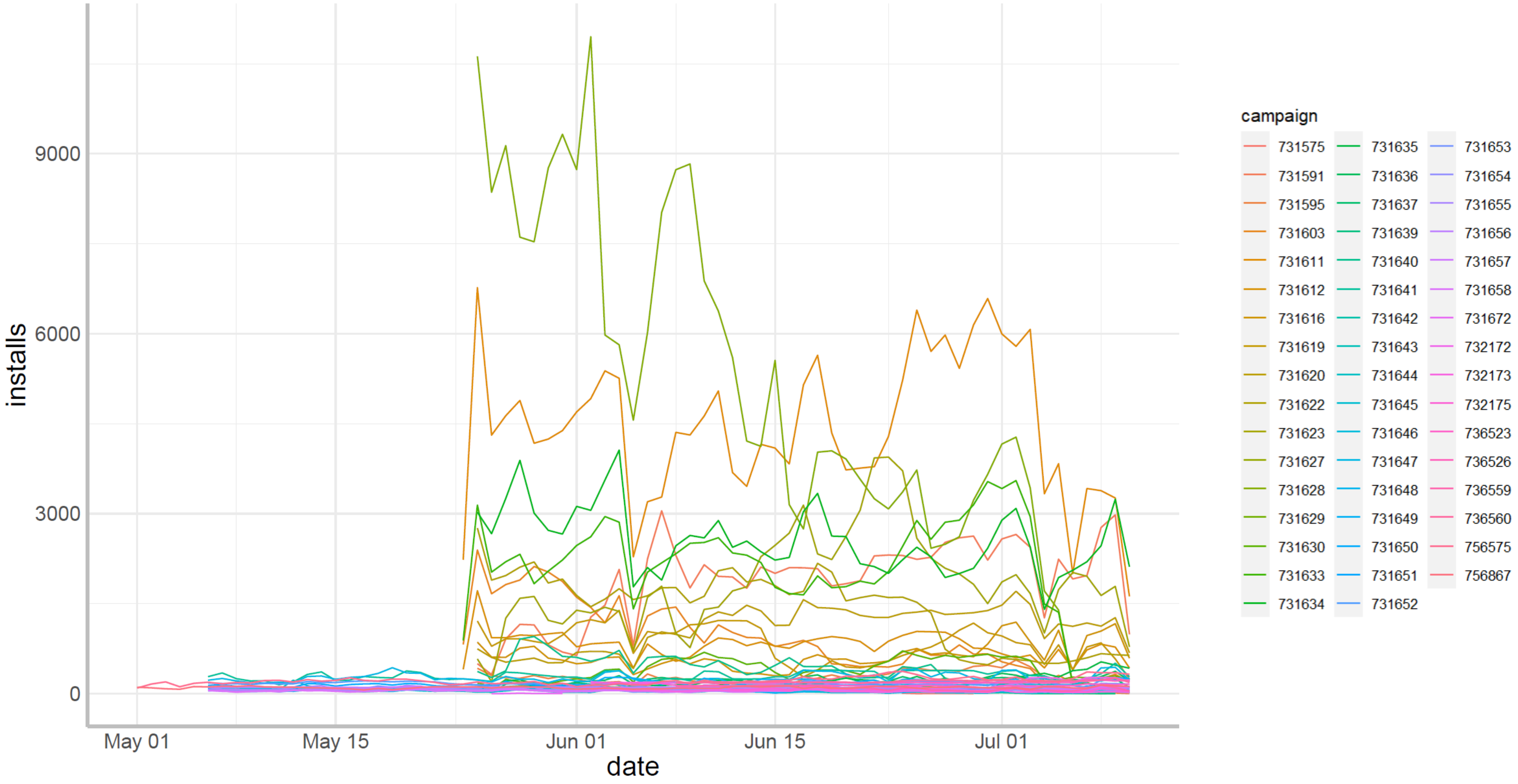
Campaign + Publisher

- It is evident that certain campaigns exclusively utilize specific publishers, and these publishers predominantly associate with those particular campaigns.



Campaign + installs

- Most Campaign install numbers are not high
- Higher installs all appear after a certain point in time



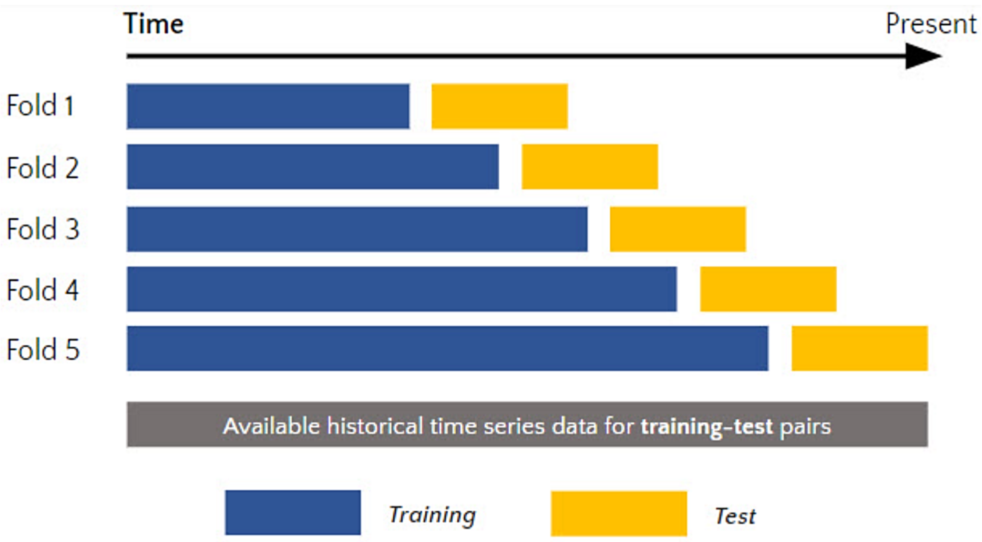
Model Building

Data Preprocessing

- **Goal:** My definition of the problem is straightforward: I am solely focused on predicting 'd90_arpu' and do not involve any of the other 'dXX_arpu' variables
- **Remove NULL:** Delete rows with empty value in 'd90_arpu' column
- **Feature Engineering:**
  - Dayofweek
  - Month
  - Dayofyear
  - dayofmonth

campaign	publisher	date	installs	d90_arpu	dayofweek	month	dayofyear	dayofmonth
731591	1	2023/5/30	333	6.327	1	5	150	30
731591	1	2023/5/31	287	3.354	2	5	151	31
731591	1	2023/6/1	267	5.127	3	6	152	1
731591	1	2023/6/2	562	4.249	4	6	153	2
731591	1	2023/6/3	652	3.127	5	6	154	3

Modeling

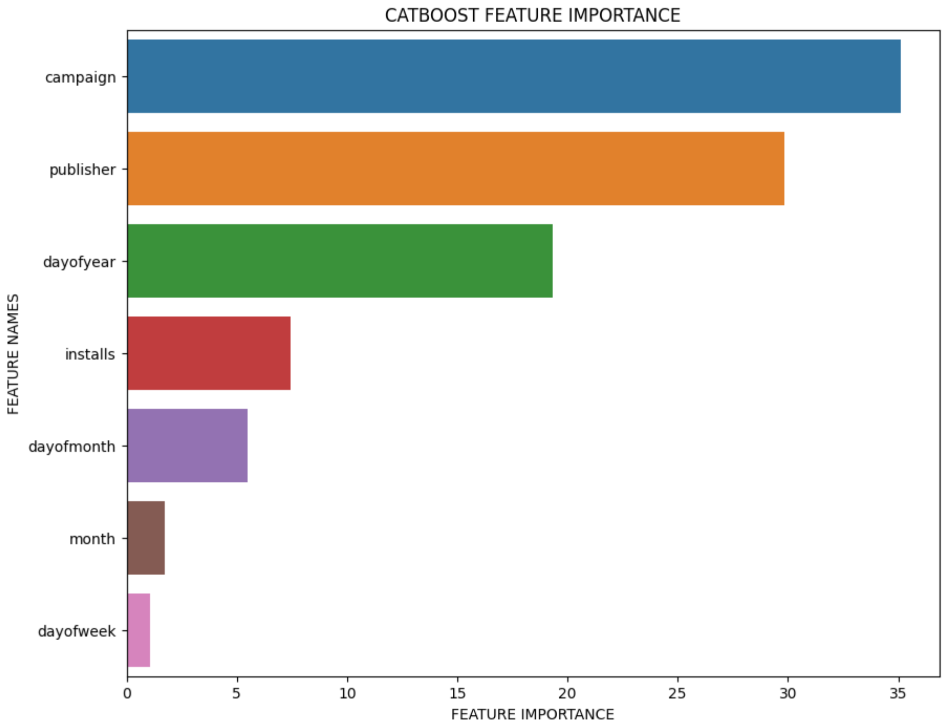


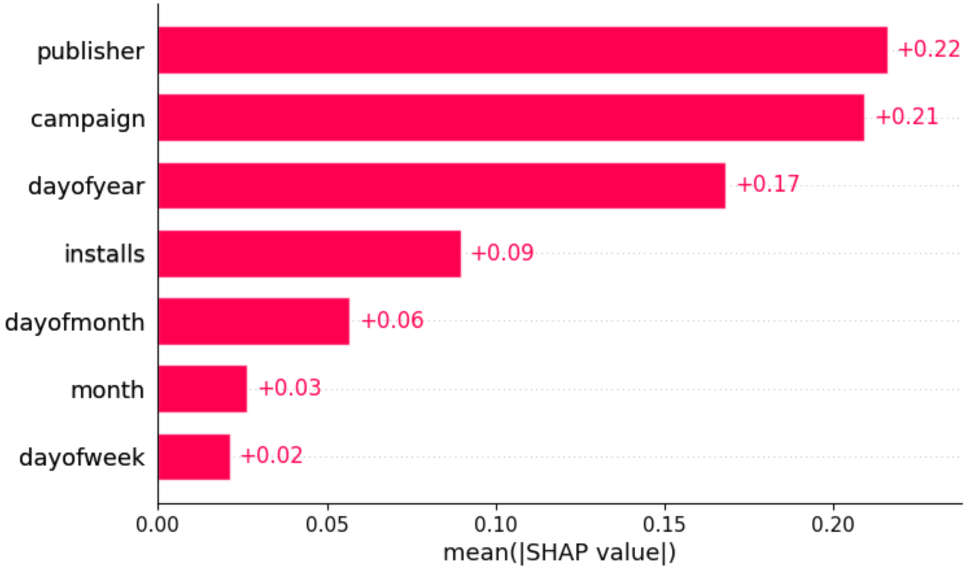
- **Validation:** Backtesting - Expanding Window
  - Fold: 3
  - Days: 10
- **Hyperparameter Tuning:** 240 combinations
  - Grid search + Backtesting: Use the hyperparameter with the lowest average MAPE in backtesting
    - Iterations
    - learning_rate
    - depth
- **Algorithm:** Catboost
  - Given the abundance of category information and the sparse data within each group, individual modeling becomes challenging.
  - Hence, I opt for CatBoost, as it can directly handle categorical data without encountering the curse of dimensionality.

Explain

Feature Importance

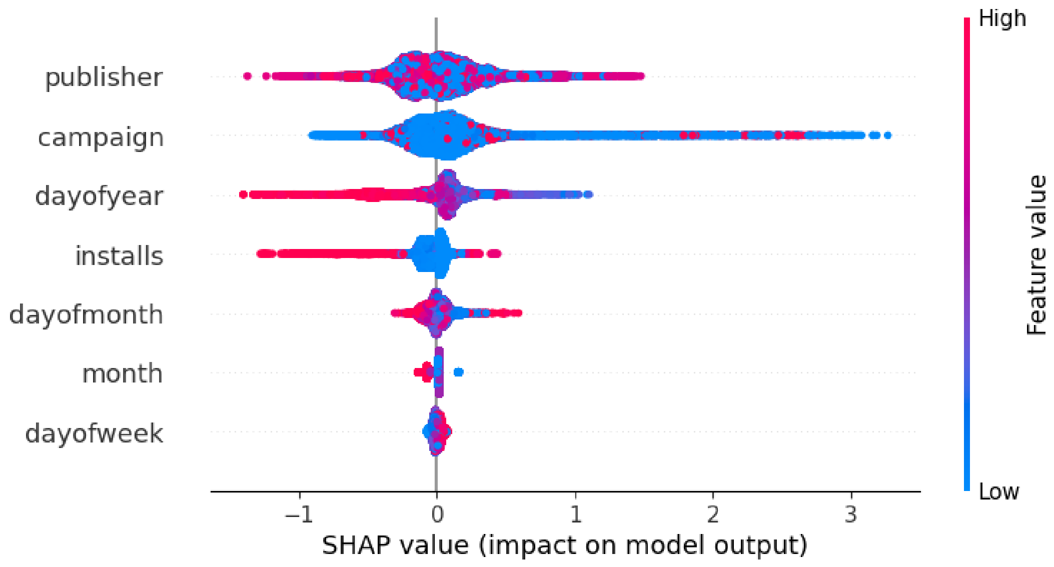
- **Tree-based Feature Importance**
  - Equal to PredictionValuesChange for non-ranking metrics and LossFunctionChange for ranking metrics
- **SHAP Global bar plot:**
  - In this plot, the global importance of each feature is determined as the average absolute value of that feature across all provided samples.
- Both assess feature importance metrics, and while there are minor distinctions, they concur that the campaign, publisher, and dayofyear columns are the top three in terms of significance.

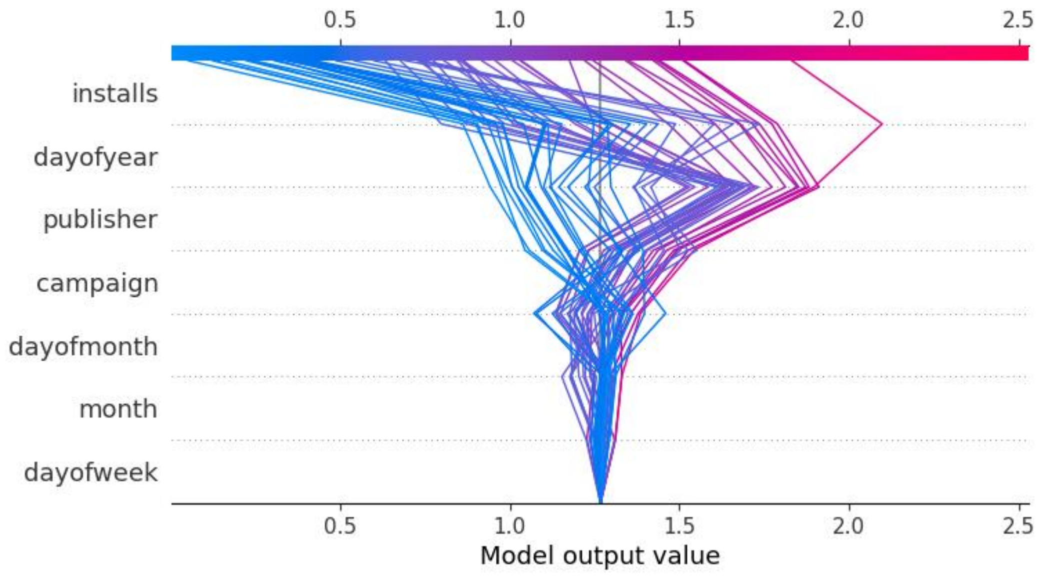




SHAP value

- **Beeswarm Plot:**
  - Absolute values of the categorical features don’t matter, because it’s hashes.
  - On the x-axis of the plot, each dot represents the SHAP value of an individual data point, offering essential insights into feature influence.
  - A broader distribution or increased density of dots signifies greater variability or a more pronounced effect on the model's predictions.
- **Decision Plot Features**:
  - Select 50 data points
  - installs is a very important deciding factor in these data points





Improvement

Optimization

- **Goal :**
  - Not enough understanding of the data, such as why null values occur. This can also explore whether information from other 'dXX_arpu' can be borrowed
- **Modeling optimization:**
  - multi-step time series forecasting
  - auto-tuning for each campaign-publisher level
  - Bayesian optimization for hyperparameter searching
- **Algorithm:**
  - Algorithm selection and optimization remain to be discussed

黃
琪 婕

CHICHIEH
HUANG

+886 956101395
cch.chichieh@gmail.com
wsxqaza12
台北, 台灣