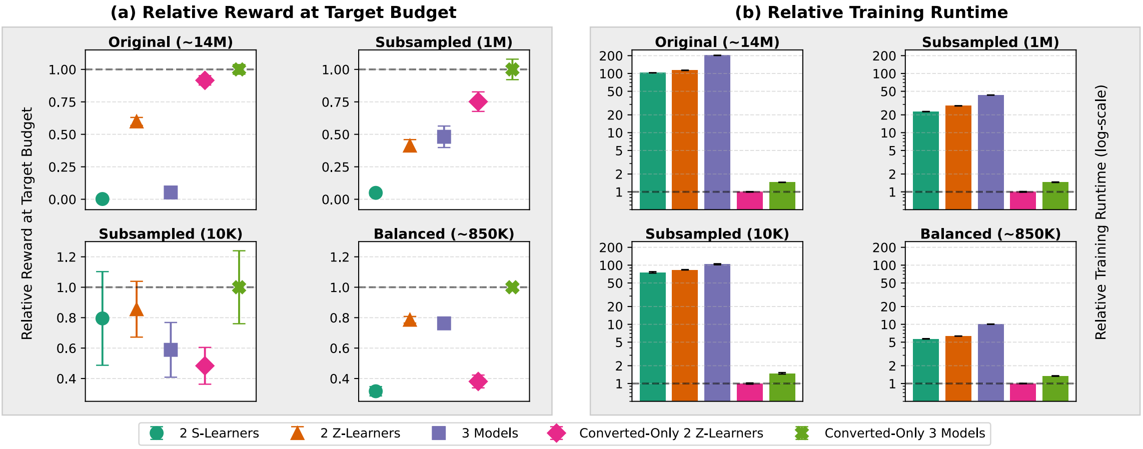

Figure 7: Experiment results on Criteo: (a) Relative reward at target budget. (b) Relative training time per method.

D PUBLIC DATASETS PREPARATION

This section describes the preparations and transformations performed on public datasets, to allow reproducibility of the paper results.

D.1 Hillstrom Dataset

Hillstrom campaign dataset [40] represents an email campaign with triggered rewards - e.g. spend (of a customer) happens only in a case of conversion. In order to simulate a cost constraint, we added a triggered cost column, such that:

$$\texttt{cost} = \texttt{spend} * \texttt{discount}$$

Where the discount was arbitrary coded by treatments: '*Womens Email*' = 10% discount, '*Mens E-Mail*' = 20% discount, and "*No-Email*" = 0%. The numerical and binary features of the dataset (`recency, history, mens, womens, newbie`) were fed to the model as-is, while the categorical features (`zip-code, channel`) were one-hot encoded.

D.2 Criteo Dataset

The **Criteo-Uplift** dataset [11] contains approximately 14 Million rows, with 11 numerical features: '`f1-f11`', and four helper fields: `treatment`, `conversion`, `visit` and `exposure`. To allow triggered cost and triggered reward setting, we have selected the fields `f2` and `f3` to represent the reward and the cost (and excluded them from the training features). These fields were chosen due to their wide distribution and low correlation between them. However, this arbitrary decision was done due to the lack of real "post-intervention" measures, and therefore, selecting other logic for reward and cost, could potentially have a significant impact on the study results.

The dataset has a conversion rate of 0.3% and only 3% of the users are exposed to the treatment. From our initial run on the full dataset, we observed that its class imbalance [35], alongside with the large-scale data and simple underlying ML-learners, lead to uninformative results in our setup (as shown in Appendix E). Therefore, we examined the results on several variations of the dataset, according to the following sampling logic:

- *Original* - Raw dataset of 14M records
- *Subsampled 1M* - a randomly sampled subset of 1 Millon records
- *Subsampled 10K* - a randomly sampled subset of 10,000 records
- *Balanced* - a balanced dataset, taking all exposed instances (420K records) in addition to the same amount (420K records) of non-exposed instances (as presented in section 5.3).

E CRITEO DATASET ANALYSIS

In Figure 7 we present the *(a) Relative reward at target budget constraint* and the *(b) Relative training time per method* on different variations of the Criteo dataset. With the full *original dataset* and its challenging class-imbalanced conversion rate of 0.3% [35], standard methods struggle with noise and treatment signals on only 3%, whereas converted-only methods demonstrate robustness. These results are limited to the specific experiment design and dataset construction - arbitrary selection of reward and cost functions ("f2" and "f3"), the underlying causal learners and ML models (S/Z-learners based on LightGBM, might struggle with log-normal features).

To validate the reason for low performance of standard models, we recreated the experiment on sub-sampled datasets. In *1 million records dataset*, we still observe a similar phenomena, where standard methods fall behind the converted-only results. In *10,000 records* datasets, the standard methods become more comparable to converted-only methods, however there is a high variance in the results due the noise introduced by random sampling. The *balanced dataset* demonstrate meaningful results of the standard approaches (2 Z-learners and 3 Models), while they still fall behind the *Converted Only 3 Models* method. From this comparison we learn that the evaluated standard models struggle to solve the problem on large or imbalanced dataset, while converted-only methods showcase a relatively consistent performance across the dataset variations.

From runtime comparison perspective, we examine the largest gap in training time of the methods in the original dataset. Reducing the training data size maintains a big gap in time between the standard and converted-only methods. However, balancing the dataset reduces the gap, since in practice it also reduces the gap in the training data size between the methods.